

INTERNATIONAL ENCYCLOPEDIA OF PUBLIC POLICY

VOLUME 4—SOCIAL, ENVIRONMENTAL AND CORPORATE GOVERNANCE

EDITOR: PHILLIP ANTHONY O'HARA

GPERU, PERTH
AUSTRALIA

2011

First published 2011
by GPERU

*GPERU is an imprint of the
Global Political Economy Research Unit*

© 2011 Editorial matter and selection, Phillip Anthony O'Hara;
Individual chapters, the contributors

Typeset in Times New Roman, Algerian, Comic Sans MS
by GPERU, Perth, Australia.

All rights reserved. No part of this book may be
commercially reprinted or reproduced or used in any other
form or by electronic, mechanical or other means, including
photocopying and recording, or any other information storage,
without permission by the publisher. Non-commercial use of
materials by individuals, libraries, universities and governments
requires proper detailed acknowledgement and statement of
access details of the encyclopedia.

British Library Cataloging-in-Publication Data
A Catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A Catalogue record for this book is available from the Library of Congress

ISBN-13: 978-0-646-97284-8

For reference only: ISBN-10: 0-646-97284-7

Libraries Australia ID: 62285981

National Library of Australia Bookmark: <https://trove.nla.gov.au/version/254124757>

INTERNATIONAL ENCYCLOPEDIA OF PUBLIC POLICY GOVERNANCE IN A GLOBAL AGE

VOLUME 4: SOCIAL, ENVIRONMENTAL AND CORPORATE GOVERNANCE

EDITOR:

Phillip O'Hara Global Political Economy Research Unit,
Curtin University, Perth, Australia

ASSOCIATE EDITOR:

Wolfram Elsner Institute for Institutional and Innovation Economics,
Bremen University, Bremen, Germany

EDITORIAL BOARD:

Anne de Bruin Massey University, New Zealand
Brian Chi-ang Lin National Chengchi University, Taiwan
Jerry Courvisanos University of Ballarat, Victoria, Australia
Wilfred Dolfsma University of Amsterdam, Netherlands
Oren M. Levin-Waldman Metropolitan College of New York, USA
Peter Muennig School of Public Health, Columbia University, USA
Jack Reardon Hamline University, USA
Celina Su City University of New York, USA
Eva E. Tsahuridu School of Management, RMIT University, Australia
Aspasia Tsaoussis ALBA Graduate Business School, Athens, Greece

How to Reference (e.g.): Kristen A. Sheeran, "Global Warming and Climate Change", in Phillip Anthony O'Hara (Ed), *International Encyclopedia of Public Policy—Governance in a Global Age, Volume 4: Social, Environmental and Corporate Governance*. GPERU: Perth, pp. 276-288, pohara.homestead.com/Encyclopedia/Volume-4.pdf

Available also from <http://independent.academia.edu/PhillipOHara>

Available also from "Z Library": <https://b-ok.org/>

Correspondence with Editor: gperu.ohara@gmail.com

Contents of Volume 4

<u>Business Ethics</u> <i>Calvin Hayes</i>	6
<u>Civil Rights, Civil Liberties & Free Speech</u> <i>Andrew Waskey</i>	21
<u>Corporate Social Responsibility</u> <i>Eva E. Tsahuridu</i>	30
<u>Counterfeiting</u> <i>Edward O'Boyle</i>	42
<u>Criminal Justice: Comparative</u> <i>Anne Cross</i>	52
<u>Criminal Justice: Punishment & Retribution</u> <i>Mark D. White</i>	63
<u>Community Health and Medicine</u> <i>Peter Muennig</i>	74
<u>Corporate Governance</u> <i>Eva E. Tsahurida</i>	89
<u>Discrimination</u> <i>Shane Ostenfeld</i>	103
<u>Economic Growth and Environment</u> <i>Oscar Alfranca</i>	110
<u>Education Policy: Distance</u> <i>James J.F. Forest</i>	123
<u>Education Policy: Preschool</u> <i>Edit Andrek</i>	131
<u>Education Policy: Schools</u> <i>Celina Su</i>	143
<u>Education Policy: Universities</u> <i>James J.F. Forest</i>	158
<u>Efficiency and Equity</u> <i>John Davis</i>	168
<u>Enterprise Net Income</u> <i>Allan Young</i>	178
<u>Environmental Governance: Community</u> <i>Anitra Nelson</i>	189
<u>Environmental Justice and Equity</u> <i>Jouni Paavola</i>	202
<u>Family Law and Family Court</u> <i>Aspasia Tsaoussis</i>	213
<u>Gender Equity</u> <i>Irene van Staveren</i>	230
<u>Geography and Governance</u> <i>Bin Zhou</i>	244
<u>Global Sex Sector</u> <i>Alys Willman-Navarro</i>	259
<u>Global Warming and Climate Change</u> <i>Kristen A. Sheeran</i>	276
<u>Green Politics</u> <i>Brian Chi-ang Lin</i>	289
<u>Health Policy</u> <i>Robert McMaster</i>	300
<u>Health and Socioeconomic Status</u> <i>Peter Muennig</i>	314
<u>Housing and Mortgage Market Governance</u> <i>Reynold Nesiba</i>	334
<u>Human Slavery</u> <i>Edward J. O'Boyle</i>	349
<u>Inequality and Distribution</u> <i>Charles M.A. Clark</i>	356
<u>Informal Economy</u> <i>Alys Willman-Navarro</i>	368
<u>Information and Communications Technology</u> <i>W. Dolfsma and F. Jaspers</i>	381

<u>Innovation Policy</u> <i>Rachel Parker</i>	392
<u>Interlocking Directorships</u> <i>Bruce Cronin</i>	400
<u>Justice, Morality and Ethics</u> <i>John Davis</i>	410
<u>Land-Use Governance</u> <i>Klaus Hubacek, Evan Fraser and Shova Thapa</i>	419
<u>Market for Corporate Control</u> <i>Ines Perez Soba Aguilar</i>	429
<u>Money Laundering</u> <i>Xiaofen Chen</i>	439
<u>Neo-Malthusianism, Population and Environment</u> <i>E.Fraser, K.Hubacek, K.Korytarova</i>	450
<u>Non-Profit Enterprises</u> <i>Robert Scott Gassler</i>	459
<u>Nuclear Energy</u> <i>Jack Reardon</i>	467
<u>Open Source Software Policy</u> <i>Alan Isaac</i>	478
<u>Ozone Layer</u> <i>Jack Reardon</i>	490
<u>Patents and Copyrights</u> <i>Wilfred Dolfsma</i>	498
<u>Pensions, Superannuation and Population</u> <i>Christian E. Weller</i>	504
<u>Pharmacoeconomics</u> <i>Edward J. O'Boyle</i>	514
<u>Pollution Rights, Taxes, Permits and Vouchers</u> <i>Jack Reardon</i>	523
<u>Principal-Agent Theory</u> <i>Andreas Feidakis</i>	532
<u>Prison Population: Ethnicity, Class and Gender</u> <i>Margaret Giles</i>	550
<u>Prostitution</u> <i>Johanna Kantola</i>	561
<u>Research and Development</u> <i>Jerry Courvisanos</i>	568
<u>Sexual Harassment in the Workplace</u> <i>Jérôme Ballet and Françoise de Bry</i>	585
<u>Small Business and Entrepreneurship Policy</u> <i>Rachel Parker</i>	597
<u>Stakeholders</u> <i>Aleksandar Sevik</i>	604
<u>Tax Evasion and Tax Avoidance</u> <i>Margit Schratzenstaller</i>	615
<u>Urban and Regional Policy Issues</u> <i>Oren M. Levin-Waldman</i>	624
<u>Welfare State</u> <i>Anne de Bruin</i>	638
<u>Wetlands Governance</u> <i>Firooza Pavri</i>	648
<u>Worker Cooperatives and Participatory Enterprises</u> <i>Roger Ashton McCain</i>	659
<u>World Trade Organisation and Environment</u> <i>Eric Neumayer</i>	674

Business Ethics

Calvin Hayes

Introduction

Business Ethics seems to combine two discordant elements, the economic imperative to maximize profits and the ethical imperative(s) to benefit (or at least not harm) societal interests. What happens when the two conflict? What happens when the two parts of the ethical imperative conflict?

While the economic imperative seems clear enough the ethical imperative is extremely contentious. For this reason the article begins with the fundamental problem of business ethics: Corporate Social Responsibility.

Corporate Social Responsibility

The problem of corporate social responsibility involves an answer to the key questions in both ethical theory and business ethics. The first, primary question is, “Who owes what to whom?” There are also some ancilliary questions, such as: “To which stakeholders is a company responsible?” and “What obligations does it have to them?” Here we should distinguish between *positive* and *negative* duties and responsibilities. A final critical question is, “Does business have a duty to make the world a better place, by for example contributing to charitable causes and hiring minorities, or is it merely obligated *not* to make the world worse?”

Answers to some of these questions also influence others. For instance, if the primary or main duty of corporations is to stockholders, this normally means that the primary duty of business is to maximize profits for shareholders. But this is always (by serious writers anyway) qualified by legal and contractual duties and, in Freidman’s case (1970), by two significant moral duties not

always legally enforceable, viz. (I) no deception or fraud and (II) following normal moral customs. The duty to shareholders is usually thought of as a fiduciary duty parallel to similar professional duties (doctors, lawyers, accountants, professors, and teachers).

For businesses that are not stockholder owned, the problem can be redefined as follows: What responsibilities (if any) does a company have to its stakeholders over and above legal and contractual obligations?

Freidman’s moral minimalism is by itself not terribly controversial. What is controversial are claims that either confine business duties to this set or expands business duties beyond this set. The latter I will designate the strong theory of corporate social responsibility. The former I will designate the weak theory.

Definitions

Every issue discussed in this article will be based on three key definitions that can be used to answer the fundamental questions of business ethics. These definitions can be used to generate and motivate all problems in business ethics whether practical or theoretical. These definitions or issues relate to “Pareto-improvement”, “corporate social responsibility” and “stakeholders”.

The Pareto-improvement criterion can be stated simply as “at least one person is better off and no-one is worse off”. In the simple version it can then apply to couch potatoes, hermits and a masochist beating herself. What is more interesting to business ethics is the corollary: “In an exchange, interaction or agreement between two or more persons or parties all are better off and no innocent third party is worse off”.

It may seem that this definition is biased in favor of utilitarian ethics. But there are two responses to this. First it is a definition of

efficiency not of ethics and second it can be redefined in “deontological” terms by replacing “worse off” with “wronged”. Further it is highly useful for defining all business ethics problems as follows: A violation of Business Ethics occurs if the Pareto-Criterion is violated. (It is a sufficient but not a necessary criterion).

Both ideally and theoretically all business transactions should meet the Pareto-improvement criterion. The fact that they all too frequently do not is what makes business ethics and /or governance necessary to control or minimize negative effects on people’s interests.

Pareto improvement can fail for two types of reasons. There may be third party effects, “negative externalities”, imposed on innocent, uninvolved third parties not involved in the original voluntary agreement. These phenomena, externalities, include pollution, the best known and most obvious. In addition there are long term marginal threshold effects that may be the result of millions of “capitalistic acts between consenting adults” (to use Nozick’s delightful phrase) that might put Mom and Dad’s coffee shop out of business or put millions out of work.

The second type of reason arises since the interaction itself may be less than ideal. This could be because one party’s consent is not completely voluntary and/or based on informed consent. This in turn entails a decision made without duress, coercion, threat, ignorance, inability to think clearly or process information, deception or fraud. If any of these conditions occur then it will not meet the first criterion. Both parties are *not* better off. In addition people change their mind after the fact. An agreement made in the cool light of reason (or without it) at one time may not seem so satisfactory later on.

There are also various criticisms of the Pareto Improvement criterion especially from the standpoint of fairness. Both capitalist and

worker may benefit from a profitable company but not equally. The same is true of producers and consumers as well as buyers and sellers.

The motive for the second definition is that almost all of the major criticisms of businesses concern allegations that they either: (1) inflict *harm* on many people unjustifiably; and/or (2) they are frequently *dishonest* and/or (3) they treat many people *unfairly*. This suggests a second criterion for Business Ethics: a violation of business ethics occurs if Common Sense Morality is violated. (This is a necessary but not sufficient criterion).

The three terms underlined constitute what will be “stipulatively” defined as “Common Sense Morality”. Common Sense Morality is founded on the view that most people would agree that harm, dishonesty and unfairness are *almost always* wrong, there are difficult borderline cases e.g. firing, fining or taxing people. But it is difficult to think of any moral criticism of business behavior not based on at least one of these three components.

The third of these requires special comment since the gist of almost any criticism of putatively unfair treatment, whether of business, government, parents, or other “authority figures” and peers, is usually the following “differential treatment of persons and/or groups *without a good reason*”. The advantage of this is that it leaves open the question: “What is a good reason for differential treatment?” This is where theories of justice become crucial, which in turn have enormous implications for the topic of governance and public policy.

The third term: “stakeholder” is the key organizing idea in the entire article. It is arguable that all problems, whether practical or theoretical can be defined with reference to stakeholders. This term, unlike the others, will be defined extensionally first then intensionally.

The following is a laundry list of stakeholders: employees, consumers, customers, competitors, stockholders, suppliers, the local community and the environment. What, if anything, do all these groups have in common? There are several ways of defining “stakeholder” intensionally but we will use the following: “Any person(s) or groups whose interests are affected positively and/or negatively by company acts, policies and decisions”.

Who are stakeholders? The definition offered earlier was intended to be both specific and clear but also neutral and therefore fluid enough to include non mainstream views (such as those to be examined under ecological ethics below).

Stakeholders are either one of the two parties involved in a putative Pareto Improvement or a third party affected and, if they are negatively impacted, it must be in terms of common sense morality. The next section will examine challenges to the entire enterprise of Business Ethics in the sense of a strong theory of Corporate Social Responsibility.

Is Business Ethics an Oxymoron, Subversive or a Pragmatic Necessity?

We will state and consider the serious objections to business ethics, not the cynical views. These are: legalistic objections, the Invisible Hand argument and, most interesting of all, the moralistic objection.

The first objection is based on the idea that the only moral obligation of business is to obey the law. Albert Carr (1968) is the patron “saint” of this movement. The key argument is that business is like a game (poker is his one and only example) where a different ethic is accepted than outside that context, i.e. bluffing is expected and therefore lying and dishonesty are acceptable contrary to a *basic principle of common sense morality*. But you still have to play within the rules of the game.

No cheating is allowed.

Given the frequent violations of law by some businesses this would be a step forward (contrary to what some critics believe). But there are ambiguities in this as well. What about producing products allowing some people to break the law such as fuzz busters or what Napster did in making it possible to violate copyright laws? Is selling guns to criminals acceptable? Interestingly Carr’s article turned out to be the most controversial in the *Harvard Business Review* up to that time. (Blodgett 1968) In addition, Carr added a gentlemanly proviso to his argument missed by both critics and supporters, i.e. that there are certain things not illegal that an ethical poker player and businessperson will not do. His approach is not as amoral as it first appears but could have been much better argued. It also ignores the point that a poker game is between consenting adults, something not always true of business, which affects many non-consenting adults and children.

Closely connected with the legalistic objection is the loyal agent argument. This holds that an employee is obligated to do whatever the firm requires because she has voluntarily contracted to be a loyal agent, i.e. to act in the interests of the firm not in either her own or society’s interests.

“Loyalty” is an ambiguous term, but on any basis it is hard to see why an employee should feel a sense of loyalty to an impersonal institution. The company will generally not feel loyalty if it feels the need to lay off employees.

It is not that hard for most people to understand why they should be loyal to family, friends, country, but to feel loyalty to IBM, Microsoft, General Motors, General Electric, is not as self-evident. Most people will agree that a man should be loyal to his spouse and children not that this duty override all other duties.

The next objection is that self-interest guided by the invisible hand leads to society being better off without any intention to do such by the actors involved.

The main problem here is due to what I term the Aristotelian proviso: viz. that many claims (especially regarding human affairs) are only “true for the most part”. The Invisible Hand works for the most part but not always, for reasons to be explained right after the final objection.

The Aristotelian proviso takes two forms in its original form (in the *Nicomachean Ethics*). They will be paraphrased here as: “Don’t expect the same precision and certainly in all academic pursuits” and “Many statements or claims are not universally true but only ‘true for the most part’”. The two fit together in the following way: A modern approach to non-universal truths is to put them in statistical and probabilistic form: (e.g. “85% of Arabs are Muslim”). The Aristotelian proviso is much less precise. When applied to the Invisible Hand this seems very realistic. Can anyone really state that the Invisible Hand works 75% of the time, or 87% or 62%? If the answer is “Yes”, then we could dispense with the Aristotelian Proviso. If the answer is: “It works 100% of the time”, then the scope of business ethics would be lowered but not altogether reduced. Smith himself spoke of the invisible hand in precisely these Aristotelian terms.

Interestingly the moralistic objection goes hand in hand with the Invisible Hand. It is a stronger argument than the loyal agent one since it rests on the legal principle of the principle/agent relation, the invisible hand, perfect competition and the Pareto-principle. But it illustrates how self-interest can undermine societal interest when the Invisible Hand is not functioning properly. This is due in part to the *principal-agent problem*.

The invisible hand argument would be decisive if it were valid. If it *always* worked

and produced Pareto improvement, then business ethics would be largely superfluous. There would still be issues of just distribution. However, it seems clear that in a system where monopoly, oligopoly and dishonest, harmful and unfair activities occur then we need to presuppose ethical behavior to make competition work.

What in general prevents the Invisible Hand from doing its beneficent work are the following: prisoners’ dilemma, moral hazard, free rider, tragedy of the commons, economic rent, adverse selection, principal-agent problem and threshold effects (many marginally harmful acts leading to massive overall harm).

Many have countered these three arguments (Solomon 1999) by advocating ethics on pragmatic grounds, i.e. more ethical companies will do better financially. Doing good results in doing well.

Both Carr and Freidman object to pragmatic arguments for business ethics on the ground that these are not moral acts, because motivated by self-interest. But whether or not intentions and motives constitute the moral good depends on one’s ethical theory. If one is utilitarian then motives are irrelevant since only results are important. So an egoist who produces the greatest good is a saint whereas the altruist, whose good intentions pave the road to hell, is not. The next section examines these contending theories.

Theories of Business Ethics

Business ethics, being intrinsically interdisciplinary makes use of a variety of theories from many disciplines. These include elementary micro-economic theory and psychological theories of moral development and cognitive psychology.

The most prominent moral theories are utilitarianism, Kant’s categorical imperative(s), social contract and human

rights theories. Recently Business Ethics has attempted to incorporate both virtue ethics (Alasdair MacIntyre in ethical theory, Solomon 1999 as an example of virtue based business ethics.) and feminist ethics as well. It also involves theories of justice especially those of Rawls and Nozick and more recently theories of “global justice”. As mentioned business ethics makes use of elementary microeconomics i.e. the theory of perfect competition and psychological theories of moral reasoning and practical reasoning. The former includes Lawrence Kohlberg and his critics (especially Carol Gilligan) and the latter psychological studies of advertising, the attitudes, beliefs and values of business leaders and consumer ability or inability to reason effectively about choices.

Lawrence Kohlberg’s (1958, 1981) theory of moral development is a good way to introduce both the more general abstract theories of ethics as well as the idea of “the expanding circle” a principle that will be examined under ecological ethics. According to Kohlberg’s theory, there is a sequence of six stages classified into three levels in the development of a person’s ability to deal with moral issues. The first level is the “preconventional”, which includes stages one and two. Usually associated with children, stage one is a concern for obedience and punishment, and stage two is a self-interest orientation. As people evolve and develop they may progress into level two, conventional behavior, usually associated with adolescents and adults. Level two includes stages three and four. Stage three links to interpersonal accord and conformity (with family and friends), while stage four is linked with authority and maintaining order (in nation or society).

Level three is the highest level of development, associated with advanced adults, including stages five and six. Stage five is a social contract orientation, while

stage six links to universal ethical principles (abstract reasoning). Stage six is associated with a critical utilitarian or deontological approach to justice, human welfare, equality of human rights and “respect for persons”. These issues are valued because of their comprehensiveness, consistency and universality, not because they happen to be the conventions of one’s society or peer group. (Kohlberg also theorized a seventh stage, transcendental morality, which was highly speculative.)

Kohlberg’s theory has been criticized on a combination of philosophical, empirical and ideological grounds. The first involves the claim that the later stages should not be assumed to be preferable. This criticism does not however dispose of the empirical claims that these are stages through which most or all infants, children, adolescents and adults move.

Carol Gilligan has criticized it from a feminist perspective, albeit one for which she claims empirical evidence. This is the argument that Kohlberg’s approach is typically “male”, focusing on impersonal, impartial and abstract moral rules, as found in principles of justice and rights, in the post-conventional stage. She argues that women are more inclined to think in terms of a caring and being responsible approach.

Business ethics, like bio-medical ethics, and professional ethics is classified as *applied* ethics. This means that the major concern is not with meta-ethical questions of interest to philosophers. It is about actually applying the standard ethical theories to practical ethical problems and dilemmas. The ones most cited in most texts are those from stages 5 and 6 and are two types: those directly relevant to practical business decisions by individuals or boards and committees in a corporation and the theories of justice which seem more relevant to governance and public policy issues. But, as with most claims in business

ethics this is also subject to the Aristotelian proviso.

Some might argue Rawls' theory of justice can be used by a company not just legislators and that Kant's universality principle should be used by governments. The claim of equal applicability is perhaps most plausibly suited for Utilitarianism. Cost benefit analysis is supposed to be applicable at both the individual and collective level and therefore applies to individuals, corporate policy and public policy.

The main problem with "utility" is defining it. It began as a very simple straightforward principle: "The greatest happiness of the greatest number" (a term associated with, albeit not invented by, Bentham). Here "happiness" is basically "pleasure minus pain". His original theory attempted to make it a purely quantitative method based on his "felicific calculus". His greatest disciple J. S. Mill dissented and introduced qualitative considerations. It has since morphed in several directions. This article will only consider the Pareto improvement criterion, and Popper's "negative utility".

Negative utility is a principle based on the asymmetry of pain and pleasure as ethical phenomena. The idea is that the demand to increase someone else's happiness, pleasure or utility has no moral claims on anyone but prescriptions to avoid the infliction of unhappiness, the prevention of suffering and the removal of unnecessary pain and suffering do make legitimate moral claims on our conduct.

The categorical imperative comes also in several versions although its defenders claim they either amount to the same thing or are corollaries of the original basic idea. The best known is the "universalizability" principle which holds that we should act on the maxim we can will as a universal law of nature *without self-contradiction*. The second version is that we should treat persons as ends

in themselves not means to an end. An extremely important point about the categorical imperative is what is termed the *reversibility corollary*: in universalizing a principle one must be willing to accept a decision about an action's ethical acceptability if the roles were reversed and one were to be victim, consumer, employee, competitor or customer rather than producer, employer, or marketer: can you agree to the ethical acceptability of torture, slavery, deception or extortion when the roles are reversed and you are the victim not the perpetrator?

To many Kant's reversibility corollary is a philosophical version of the golden rule: do not do to another person what you would not wish done to you. However, Kant argued for this on rationalist grounds, not merely in the sense that it is not based on religion or revelation but is based on *pure* reason i.e. it is not empirical in the sense that one would not *want* to be tortured, enslaved, beaten, used like a machine or tool or robbed, defrauded and deceived. One cannot *logically* will this. It would be self-contradictory Kant argues to will theft as a universal law of nature for the following reasons: one must simultaneously will the institution of property (or there will be nothing to steal) and will its non-existence (or I cannot steal it).

The underlined terms are crucial in distinguishing Kant's original version from later attempts by utilitarians (from J.S. Mill to R.M. Hare) to co-opt this principle for their use. Kant's concept is rationalistic and deontological whereas utility is an empirical principle. Where it is supposed to make a difference is in (at least) two main areas: promise-keeping and considerations of justice (both retributive and distributive). It is therefore striking that the two major theories of justice from the 1970s those of Rawls and Nozick are both explicitly anti-utilitarian.

Before we look at them another crucial

distinction needs to be clarified. This is the distinction between *positive* and *negative* duties whether we are dealing with rights, utility or the golden rule. (Its' similarity to the reversibility corollary has already been noted).

Negative rights are essentially defined by containing the word "Not". They include the right not to be harmed, or violated by assault, murder, theft, *inter alia*. Positive rights entail a duty to act and not merely refrain from harming others: prime examples are putative rights to health care, education, welfare and adequate housing. Similarly the golden rule can be stated with or without a "Not": "Do unto others what you would want them to do unto you".

At the risk of oversimplification the main difference between Rawls and Nozick is that the latter emphasized negative rights and duties; the former both.

Rawls' theory (1971) goes well beyond the negative advocating distributive or "social" justice. Since his equal liberty principle is not as controversial as the difference principle I will focus on the latter. It holds that differences in wealth, income, power and privileges in society are justified if and only if they are subject to the first principle and work out for the benefit of the least well off.

Nozick (1974) attacks this with his own entitlement theory. It holds that it is the *process* not the *product* or outcome that is just or not. If property is both legitimately acquired (as in a Lockean state of nature where all property is un-owned) and then legitimately transferred it is a just process. There is also a proviso that there must be rectification for illegitimate acquisitions and transfers (fraud, force, theft, forcibly excluding others, violence).

While it may seem that the implications are straightforward for governance and public policy this may not be the case. Nor are the implications so straightforward to one of the

hottest topics today: intellectual property rights. This point will be argued the final section.

But is it reasonable to expect business persons, with conflicting demands and heavy schedules, to take time out to read Kant's *Metaphysics of Morals*, or busy legislators and civil servants to read John Rawls' *A Theory of Justice* and then turn them into practical politics for both business and policy governance?

Can these ideas be turned into more down to earth principles? The term "respect for persons" can be easily translated as "don't treat employees or consumers as things: tools or machines, to be discarded, sold or traded when no longer of use". This does not per se rule out firing or demoting persons but it makes the key point: people are not things to be used as means to an end, including profit or overall societal utility.

The *universalizability principle* has, as one reasonably clear implication, that any type of double standard, whether applied to gender, race or any other basis for discrimination cannot be justified without very good reasons. Since it is closely related logically to the criterion of Reversibility it is sufficient to note the already made point that it is a more abstract, philosophical variant of the golden rule. Could you accept the principle applied in this case if the roles were reversed?

The Pareto-criterion is equivalent to the idea of a win-win situation for both sides unlike the zero-sum game or negative sum game of the prisoners' dilemma.

Critical Issues of Business Ethics

The first list of stakeholder problems concerns the most obvious and perhaps important stakeholder: employees. Some of the major issues are: health and safety, whistle-blowing, privacy, pay equity and employment equity, employment at will vs. due process, participatory management,

“diversity” and problem arising from layoffs, plant closings, part-time employees and “meaningful” work.

All of these can be defined with the three basic principles of common sense morality but only three examples will be used for reasons of space.

Whistle-blowing is a problem because it seems to involve a conflict between an employee’s duty to the company to be a loyal agent and her duty to be a responsible citizen and so report wrong doing especially that which is illegal and likely to be extremely harmful to innocent citizens.

Pay equity and employment need to be distinguished from two traditional concepts they are often confused with. Pay equity goes beyond *equal pay for equal work* to include equal pay for work of *comparable value* or worth which requires a complex, quantitative assessment of differing jobs: those that are male dominated and those that are female dominated. Employment equity goes beyond equal opportunity, calling for positive action on behalf of previously victimized minorities: in North America especially Afro-Americans, First Nations and other visible minorities.

Employment at will is a key issue in public policy debates. It involves the simple principle that employment exists at the mutual consent of both parties and so each party must agree to the original contract (ruling out slavery or serfdom) but also either side can terminate at their will: so the employee can quit but he can also be fired with or without a good reason, or any reason for that matter.

Issues of gender/sex include the glass ceiling and sexual harassment. While this list obviously overlaps with employee problems they are worth stating separately. One of the striking features of employee problems concerns the difference faced by male and female employees. As a general rule the former concern issues of harm (especially

health and safety) whereas the latter are issues of justice or fairness. The vast majority of injuries and deaths on the job are male victims whereas female complaints concern pay employment and harassment issues (Warren Farrell 1993).

The major topics under consumer rights issues involve product safety and reliability, the effects of marketing sales and advertising on consumer autonomy, honesty and terms such as “informed consent”.

In many ways it is arguable that the main issue in advertising ethics is or ought to be honesty. The dependence effect (argued by J.K. Galbraith 1958) is dubious and contestable. It was effectively criticized by Hayek (1961) and recently by Michael Phillips (in Shaw 2003). Other problems: bad taste, insulting commercials and hiring of celebrities to hype products are less ethical than etiquette or aesthetic issues.

Under ecological problems it is crucial to distinguish four main types: pollution, resource depletion, future generations and animal rights. It raises the tricky issue of whether unborn people or animals (including species) can be regarded as stakeholders. In this area, utilitarianism can claim an advantage over Kant’s Categorical Imperative. This is because pain and suffering can be experienced by non-humans and non-persons.

Here is where two of the less elegant terms in business and ecological ethics: “speciesism” and “considerability” emerge. The former is deliberately analogous to “sexism” and “racism” and related to the expanding circle argument. The latter is a concept raising the question: “Whom (or even what) should we *consider* when we are considering the impact of our actions, policies and decisions on the environment?” This includes business acts and decisions and public policy as well.

There are five main candidates. The first is

anthropocentrism; where only humans count either in terms of duties or rights (although most versions condemn cruelty to animals). The second is sentientism; where all and only sentient creatures (those capable of experiencing pain) are entitled to moral consideration. The third is bio-centrism; where all life is entitled to moral consideration, including trees and butterflies, snail darters and virus carrying insects. The fourth is where eco-systems are (or ought to be) the primary target of our ecological concern (e.g., rain forests). And the fifth candidate is deep ecology; where any and all existence should be considered sacred and not subject to humanity's control.

The expanding circle argument begins by assuming a successful transition from Level 2 (which entails stages 3 and 4) in Kohlberg (loyalty to family and friends, stage 3, loyalty to nation, stage 4) to Level 3, a universalist, cosmopolitan outlook and then urges that we must move beyond this. The main criticism to be made against the expanding circle argument is that it assumes facts not in evidence (and unargued principles). It assumes we have somehow persuaded the vast majority of the 6 billion people in the world to reach stages 5 and 6 in Kohlberg's hierarchy and now must invent stages 7 and 8 for them to advance up to: first to include all life (stage 7) and then all existence (stage 8). What seems to be true, instead, is that the vast majority of people remain at Level 2 and never ascends to Level 3.

When it comes to issues of globalization and business investment in Newly Industrializing Countries there are a host of practical issues under girded by several theoretical problems ranging from debates about theories of universal human rights and global justice as well as cultural relativism.

Cultural Relativism rests its claims on a set of premises, which are factual, epistemic, logical and even metaphysical whenever it

merely ceases to be a dogmatic mantra. It requires more than the (by now) rather trite observation that ethical beliefs and values vary both diachronically and synchronically. It rests at its best on the Weber-Robbins position that facts and values (or "is" and "ought") are both epistemologically distinct and logically incongruent. This means factual propositions can be proven or disproved (empirically or logically) whereas value judgments or moral principles cannot be and that the latter cannot be logically deduced from the other.

It is at least arguable that the real problem is not theoretical but pragmatic. Consider the following: Do torture, corruption, oppression of women and suppression of dissent exist in such countries because the people in those societies value then (unlike Westerners) or do they exist because those in power are able to get away with them?

The major religions in Asia are Hinduism, Buddhism, Confucianism and Islam, but where exactly do their sacred texts express approval of bribery and extortion? Have any of the many spokespersons for the superiority of "Asian Values" included corruption among these values? Cultural relativism, while often used in a misguided way to encourage "tolerance", diversity and multi-culturalism, gives both business *and the state* convenient excuses to abuse people, in a manner that universal rights cannot do so.

Business Ethics in the 21st Century

The major problems for both corporate governance and public policy that society and business will face in the (early) 21st century arise from the following sources: biotechnology, information technology, globalization, intellectual property rights, poverty and equality, modernity and westernization and debates about universal human rights.

Consider the duo, westernization and

modernization. Do the two have to go together? While there is little doubt that the vast majority of people in the non-western world want the benefits of the latter (better health, more wealth, increased longevity) there can also be little doubt that many would prefer it without what is often perceived as the downside of westernization. Radical Islam is only the most obvious and extreme instantiation of this attitude.

Fundamentalist Muslims and “socially conservative” spokespersons for Asian values are not adverse to modernity but they are opposed to the “moral decadence” of the west (as are a significant minority of westerners).

While some of the issues are not new (privacy, poverty, eugenics) their nature is changing while others clearly are new (cloning, stem cell research, genetic screening). Frozen embryos raise questions related to, yet quite distinct from, the problem of abortion. In many ways the most interesting issues in the early 21st century concern the overlap between biotechnology, information technology, ecology and globalization.

Cloning is a problem because many see it as an attack on “human dignity” or perhaps even our definition of “human being” or “human nature”. It is easy to see, at least theoretically, the benefits of cloning animals—it should result in increased production (not just of meat) at cheaper cost, due to less scarcity. Of course the costs of cloning have to be factored in but once the usual economies of scale are achieved then it should have the usual beneficial results. However, “What are the benefits of cloning humans”? In reply it can be asked, “What are the harms”? Apart from the usual scary scenarios (cloning Adolf Hitler or Charles Manson) there are two primary objections: first if the world is as seriously over-populated as may claim, then why add to the problem? Second, and more basic: what if the

cloning is botched and we produce a “monster”: a human with four feet or three arms or two heads? The experience with animal cloning is instructive here, since it took numerous attempts to produce “Dolly”, the first cloned sheep.

Given these arguments it is reasonable to suggest that the best public policy would not be an outright ban but a moratorium for, say, at least ten years. Then it could be reconsidered after several years of both experimental research and very careful public and expert debate.

Stem cell research is a problem because, while it promises great benefits, curing diseases being the most prominent, it also seems too many to involve the deliberate creation of at least a proto-human being and then destroying it. In addition it is an open question whether the same benefits could be achieved in other less contentious ways such as via adult stem cells.

Fertility clinics present totally unprecedented problems for both business and public policy. While they promise help for couples unable to procreate, they also involve the use of frozen embryos and the problems of surrogate motherhood. A very interesting problem for the standpoint of business ethics and public policy is whether altruistic surrogacy is acceptable but not commercial surrogacy. Is the latter the same as “buying babies” or is it merely providing a service at a reasonable monetary cost? The former is illustrated by a woman who is a surrogate mother for her infertile sister but charges no fee (over and above incidental expenses perhaps).

What happens if a couple contributes frozen embryos to a clinic and then divorce? Who has “ownership” rights? Are embryos property, persons or neither? What if a couple engage the services of a surrogate mother and she changes her mind and wants to keep the baby? (This is not a purely hypothetical

example: it has happened.)

The problem of eugenics can be clarified both logically and historically by the following two points: It can be usefully compared logically to both modern hopes (and fears) about genetic therapy and genetic enhancement. Historically it is worth noting that in its heyday in the 1920s, it attracted not just “right-wingers” such as Hitler (who no doubt is most responsible for its thorough discreditation), but also progressive thinkers such as Margaret Sanger (a feminist hero for her early advocacy of birth control), Julian Huxley and a significant number of American progressives and their European counterparts.

Genetic therapy is the attempt to make use of our rapidly increasing knowledge of genetics to rid the world of the terrible diseases many children inherit at birth. This would seem to most people to be eminently desirable, especially to the parents involved. Genetic enhancement on the other hand refers to the use of the same type of knowledge to improve a child’s characteristic which are not debilitating but which, if improved may improve their chances in competition for society’s rewards. If being taller, stronger, having a higher IQ or possessing greater musical or athletic talent gives a child an advantage and is sought through genetic engineering then it is enhancement not therapy. It is basically the difference between avoiding something undesirable and achieving something desirable.

Eugenics and genetic screening raise difficult dilemmas due to recent developments in biotechnology that will give both businesses and makers of public policy much to ponder. The possibility of determining either with certainty or high probability the defects a child may have while still in the maternal womb make it possible to use abortion as a method of eugenics. What if many parents decide to abort a child they know she will be a Down’s Syndrome child

or have a low IQ but would not make the same choice other wise? Is this eugenics or pro-choice in action?

Genetic screening is a procedure that many businesses including insurance companies might wish to be able to do. A company may, for either humanitarian or self-interested reasons, wish to keep higher risk employees away from harmful chemicals. This would simultaneously reduce costs but also be regarded by many as a violation of privacy and as unjustified paternalism.

It also raises issues of fairness in the case of insurance companies, a dilemma no matter which side one is on. It would seem unfair that a person whose innate genetic defects are clearly not her fault should be the victim of discrimination (either being uninsurable or paying higher premiums). But is it fair that the insurance companies, who are equally innocent, pay the extra costs? But can the insurance companies acquire this information without violating rights to privacy?

This problem provides a link to the next—the use of information technology for purposes that can be beneficial, harmful and lie in a gray area. The problem of genetic screening and eugenics illustrate the intimate connection, since it is the ability of computers to access and rapidly calculate numerical data that make *Brave New World* scenarios seem all too realistic. In the movie *Gattaca*, an early scene shows a needle connected to a computer inserted into a newly born infant’s foot. The computer then grinds out reams of paper with calculations of the probability of several different types of diseases and health problems in his future life.

As computers make it easier to store reams of data on numerous people to what uses, misuses and abuses can these be put? As technology makes it easier to copy others’ music, books and tapes should government continue to enforce patents and copyrights or do these laws need to change to reflect the

“new realities”? One of the more interesting problems here is that it seems (ironically perhaps) that the ideology most genial to the free market, libertarianism, is least congenial to the whole idea of intellectual property rights.

It is not at all obvious that they can be justified on any of the standard rationales for property rights. This is because of two factors: first, the scarcity rationale is (or appears to be) less compelling. If I copy your music, book or disc you still have the original contrary to the situation where I steal your car, burn your home or vandalize your store. It also restricts my use of my own property legitimately acquired.

In a manner similar to the Wilt Chamberlain extra income example, (Nozick 1974) a person downloading music on her own computer or taping music, movies and TV programs with tapes she has purchased on machines she owns in a impeccably libertarian, Nozickian fashion has not engaged in acts such as stealing, defrauding, enslaving or “seizing other’s products and preventing them from living as they choose, or forcibly [excluding] others from competing in exchanges” (Nozick 1974). As a matter of fact (or logic) it seems that companies relying on the law to prosecute such “piracy” are doing the last named activity. This should be distinguished from blatant plagiarism, stealing confidential information and perhaps, from patent laws. Most justifications of patents and copyright are utilitarian- without them there would be little incentive to innovate and produce new desirable products whether medical, musical or whatever.

This topic cannot be given justice here so I will finish with two points. First it seems obvious (if anything is in these contentious areas) that incentives for innovations in medical products are more desirable and urgent than those in music. Second, Lester Thurow (1997) has suggested some major

changes in intellectual property law to take into consideration significant differences in types of inventions that can be patented and opposes the “one size fits all” mentality suggesting many changes in order to combine the advantages of both efficiency and fairness.

The *precautionary principle* is extremely interesting and controversial, especially in connection with biotechnology. Lacking any evidence that a technological innovation can (or may probably) cause harm, yet uncertain whether it *may*, (either in the short run or in the long run) should we permit it or require more testing? In the extreme version it holds that the burden or proof is not on the critic who suspects possible harm but on the defender who may be able to point to demonstrable benefits but cannot prove there are not any, as yet undiscovered, harms. The strong version holds that “a product should not ever be marketed unless it can be shown that it is *not* harmful”.

Curiously this principle (often associated with left wing critiques of biotechnology) is both hyper-Burkean and Luddite as well, very strange positions for left-wing “progressives” to adopt. The more significant objection however is that this principle would shift the burden of proof in an impossible sense and would also have prevented most of the beneficial inventions and innovations of the past four or five centuries from occurring.

There are many who argue the potential of biotechnology (Michael Fumento, Gregory Stock) to solve or at least significantly contribute simultaneously too many of the greatest problems facing the world in the early 21st century: it promises huge gains in agricultural productivity, the curing of terrible diseases, indefinite prolongation of the average human lifespan and the minimization of ecological degradation. The benefits, it is also argued, will not be confined to the already affluent but be widespread among

farmers and consumers in Africa, Asia and the poorest people everywhere, thus meeting Rawls' difference principle desiderata.

This is one area where public policy debates will be extremely difficult for the legislators but also where a cost-benefit analysis should be of great value. This can be done and still combine fairness, not just utility in the solution since cost-benefit analysis should primarily be a method of analysis and not the sole criterion of choice, or at least not without further argument.

If the optimists turn out to be correct, then biotechnology may make recent heated debates about globalization and its relation to poverty and inequality irrelevant. There is however one caveat that must be entered here. Discussions of poverty and inequality often conflate the two problems and thus beg several questions. Yet the two are clearly distinct. Consider the possibility of two societies in which the following scenarios hold. In Society E, everyone is very equal or almost entirely so, but everyone is below the poverty line (however defined or measured). Then consider Society G, where everyone is above the poverty line (same proviso as above) yet the Gini coefficient is very high. Society E has "solved" the problem of equality but not poverty whereas Society G has solved the poverty problem but not that of equality.

A separate examination of one possible benefit within the discussion of Fumento and Stock is appropriate. This is the possibility of a very huge increase in human life span, not necessarily as far as that of the legendary Methusaleh, but a doubling or increase to 150-200 years, not implausible scenarios according to some experts on such research. Would such an increase be desirable? It is not merely the quality of life factor here: "Would these people merely be surviving an extra 70-100 years on life support systems with no ability to engage in typical human activities?

Would they be like the immortals in *Gulliver's Travels* who merely age and decay at the same rate?"

Could the world cope with such a phenomenon? If people continue to reproduce even at today's reduced rates could we cope? How high would the retirement age have to be raised? Would this be the ultimate Malthusian nightmare?

Whatever the result of these debates, we have in effect returned full circle to the problem in paragraph 1. How do we balance the undoubted benefits innovative businesses and entrepreneurs can bring society against both the uncertain risks and actual harms that they also bring? How can we minimize the latter, maximize the former and do both in a manner that is as fair and impartial to all of humanity (and perhaps all sentient creatures) as is possible?

I will conclude with a short argument in favor of a position between the weak and strong theory of corporate social responsibility. The Invisible Hand needs both Virtue and Law on its side. The former will make society better while the other two are needed to prevent or minimize possible negative third party effects. It is therefore arguable that the achievement of getting business, politicians, ordinary citizens and whoever else to scrupulously respect only the negative duties of corporate social responsibility would be the closest we will or ever can come to Utopia.

The implementation of positive rights then would be left to more appropriate institutions: family, religious and eleemosynary institutions. When businesses do contribute to charity it should be treated arguably as desirable but not obligatory: what used to be called supererogatory acts, i.e. those going above and beyond the call of duty. A striking example of this is Merck's contribution of many medicines either free or below cost in Africa, especially to help cure the terrible

disease of river blindness. This has the added advantage of not expecting businesspersons to be saints while praising them when they occasionally act ethnically. This too has a parallel advantage of not expecting politicians (or their “expert” advisors) to be angels or gods. In addition if they can master the simple principles of common sense morality they need not be sophisticated philosophers either, as long as they can think and reason logically and critically.

Selected References

- Blodgett, Timothy B. (1968) “Showdown on Business Bluffing”, *Harvard Business Review*, Volume 46, Number 3, May/June, pp. 162-170.
- Carr, Albert. (1968) “Is Business Bluffing Ethical?” *Harvard Business Review*, Volume 46, Number 1, Jan./Feb, pp. 143-153.
- Crampton, Peter and Dees Gregory. (1993) “Promoting Honesty in Negotiation: An Exercise in Practical Ethics”. *Business Ethics Quarterly*, Volume 3, Number 4, pp. 359-394.
- Dodds, Susan; Lucy Frost; Robert Pargetter; and Elizabeth Prior. (1988) “Sexual Harassment”, *Social Theory and Practice*, Volume 14, Number 2, pp. 111-130.
- Friedman, Milton. (1970) “The Social Responsibility of Business is to Increase its Profits”, *New York Times Magazine*, 13 Sept, 122-126. Reprinted in J. Des Jardins and J. McCall. (1985) (Editors), *Contemporary Issues in Business Ethics*. Belmont: Wadsworth, pp. 21-25.
- Fumento, Michael. (2003) *Bio Evolution: How Biotechnology is Changing our World*. San Francisco: Encounter Books.
- Galbraith, J.K. (1958) *The Affluent Society*. Boston: Houghton Mifflin.
- Hayek, F.A. von. (1961) “The Non Sequitur of the ‘Dependence Effect’”, *Southern Economic Journal*, Volume 27, Number 4, April, pp. 346-348.
- Farrell, Warren. (1993) *The Myth of Male Power*. New York: Random House.
- Frank, Robert. (1999) “Can Socially Responsible Firms Survive in a Competitive Environment?”, in Thomas Donaldson and Patricia Werhane (Editors), *Ethical Issues in Business*. Sixth Edition. New York: Prentice-Hall.
- Hosmer, LaRue Tone. (2003) *The Ethics of Management*. Fourth Edition. New York: McGraw Hill Irwin.
- Kohlberg, Lawrence (1958). *The Development of Modes of Thinking and Choices in Years 10 to 16*. PhD Dissertation, University of Chicago.
- Kohlberg, Lawrence. (1981) *Essays on Moral Development, Vol. I: The Philosophy of Moral Development*. New York: Harper & Row.
- Nozick, Robert. (1974) *Anarchy, State and Utopia*. Basic Books, 1974.
- Rawls, John. (1971) *A Theory of Justice*. Harvard: Harvard University Press.
- Schwartz, Felice. (1989) “Management Women and the New Facts of Life”, *Harvard Business Review*, Volume 67, Number 1, Jan/Feb, pp. 65-76.
- Shaw, William H. (2003) *Ethics at Work: Basic Readings in Business Ethics*. Oxford: Oxford University Press..
- Sen, Amartya. (2002) “Does Business Ethics Make Economic Sense?”, in Thomas Donaldson and Patricia Werhane (Editors), *Ethical Issues in Business: A Philosophical Approach*. Seventh Edition. Upper Saddle River, NJ: Prentice-Hall, pp. 244-255.
- Singer, Lysonski and David Hayes. (1991) “Ethical Myopia: The Case of ‘Framing’ by Framing”, *Journal of Business Ethics*, Volume 10, pp. 29-36.
- Solomon, Robert. (1999) *A Better Way to Think about Business: How Personal Integrity Leads to Corporate Success*.

Oxford: Oxford University Press.

Stock, Gregory. (2002) *Redesigning Humans: Our Inevitable Genetic Future*. Boston: Houghton Mifflin.

Tapscott, Don and Ticoll, David. (2003) *The Naked Corporation*. Vancouver: Viking Press.

Thurrow, Lester. (1997) "Needed: a New System of Intellectual Property Rights", *Harvard Business Review*, Volume 75, Number 5, Sep/Oct., pp. 94-103.

Velasquez, Manual. (2002) *Business Ethics: Concepts and Cases*. Fifth Edition. New York: Prentice Hall.

Zimmerman, Michael, (1998) (Editor) *Environmental Philosophy: From Animal Rights to Radical Ecology*. Second Edition. New York: Prentice-Hall.

www.yahoo.com/Government/Law.

Calvin Hayes
 Faculty of Business
 Brock University, Canada
 chayes@brocku.ca

Websites

Anti-Trust Organization. www.antitrust.org.

Business Ethics Organization.
www.businessethics.org

Business Social Responsibility Organization.
www.bsr.org

Corporate Watch. www.corpwatch.org

Envirolink. www.envirolink.com.

Environmental Fund.
www.efund.com/investors_action

Federal Trade Commission. www.ftc.gov.

Greenmoney Fund. www.greenmoney.com

Hieros Gamos. www.hg.org.

Multinational Monitor. www.essential.org

Occupational Health and Safety Association.
www.osha.gov.

Pacific Net Page. www.pacific.net.

Santa Clara University.
www.scu.edu/SCU/Centers/Ethics

Students for Responsible Business.
www.srbnet.org

Wall Street Research Net. www.wsrn.com

Website for Manual Velasquez.
www.prenhall.com/velasquez

WorldWatch. www.worldwatch.org

Yahoo Anti-trust Links.

Civil Rights, Civil Liberties & Free Speech

Andrew Waskey

Introduction

Civil rights and *civil liberties* are terms that are often used interchangeably. However, there are very significant differences between the two terms.

Essentially civil liberties are the rights that individuals have in opposition to the power of governments to regulate their lives. Civil liberties limit governmental authority and create rights that the government is obligated not to violate. In contrast civil rights limit the actions of private persons whether the private persons are an individual, a group of individuals, a corporation acting as a private person, or even the government itself acting after the manner of a private person or corporation.

Civil rights are protected by governments acting through their legal and political powers as the protector of individuals, or even groups or classes of people, who may be denied equality by private parties, or even by the government itself. Affirmative action programs have been instituted in a number of countries to end discrimination that is now believed to be a denial of claims to equal treatment in civil life. That is the denial of civil rights by private parties can be prosecuted by a government in the interest of a policy goal of equality between persons regardless of race, religion, color, gender, or other arbitrary criterion.

Governments can be the creators of civil rights. When a government enacts legislation such as a veteran's benefits or retirement benefits program it is creating civil rights. It may even create them as benefits to educational opportunities through access to credit to pay for education (student loans). These positive civil rights are entitlements that governments give, usually administer, and can

end if it is found to be expedient to no longer maintain them.

Historically civil liberties have been rights such as, free exercise of religious practice, freedom of speech, press, assembly, to petition the government for the redress of grievances, to be free from abuse by the government in criminal investigations from unreasonable searches and seizures, or not to be tortured into confessions, or not to be victimized by double jeopardy, denial of legal counsel, not to be subjected to cruel and unusual punishments, or similar rights. These are rights that a citizen needs in order to be able to criticize the government, to join with other people spontaneously or as an organized group in order to criticize the government or to present petitions for the redress of grievances.

In a restricted sense criticism of the government can be seen as a threat to it. In the American political experience in the early history of the Republic the Federalist Party headed by President John Adams adopted the Alien and Seditions Acts (1798) which effectively made criticism of the government seditious. The Acts were so alarming that they led to a crushing defeat of the Federalist Party in the Election of 1800.

What the Adams Administration failed to distinguish was that criticism of government policy and actions to overthrow the government are not the same. The idea of the loyal opposition was a lesson that the Federalists learned at a great political price.

In effect the greatest power of a citizen is not the right of freedom of speech, but the power of the ballot cast in the ballot box. The power to criticize whether fairly spoken or even poorly written is important but greater is the right to vote. The vote gives people the power to cast the rascals out of office.

Civil rights, in contrast, can be defined as the rights that individuals have in relation to other individuals. Civil rights are also those positive acts of governments that ensure that someone

is treated as an equal member of society. It would include punishing racial or gender discrimination.

For an individual to have civil liberties implies a theory of limited government. It means that individuals have an area of freedom that can not be easily invaded by the government. It is in recognition of this limitation that the Constitution of the United States of America has a Bill of Rights. Originally it was intended to limit the Congress of the United States, but in the Twentieth Century the United States Supreme Court "selectively incorporated" or nationalized much, but not all, of the Bill of Rights to make it applicable to the American states as well. The Bill of Rights has also been extended to applied to the whole of the federal government and not just to the Congress.

For an individual to have civil right implies a theory of government exercising its power on behalf of someone to protect liberties from discriminatory treatment. Civil liberties include access to public spaces. When the American Congress passed the Civil Rights Act of 1964 it outlawed racial discrimination in places of public accommodation, thereby opening the doors of restaurants, hotels, and other public places to racial minorities for equal treatment.

Development of Civil Liberties and Rights

By the beginning of the 21st Century many rights had been codified in domestic and international codes. These liberties and rights are products of many historic struggles in Europe and later in America over issues such as freedom of conscience in matters of religion, freedom to engage in commercial activities, or the exercise of historic rights.

The English colonists who immigrated to the original thirteen American colonies believed that they inherently carried with them the "rights of Englishmen". From the Magna Carter to the American Revolution the idea that people had rights that could not be

violated legitimately by anyone, including the monarchy or nobility, grew and spread. Doctrines of natural law developed from the ancient Stoics and of natural right developed, especially by social contract theorists in the eighteenth century, strengthened the foundation of modern civil liberties formed the foundation for modern ideas of civil rights and civil liberties.

However, in the Nineteenth Century the advancing authority of science was used by a number of forces to reject the claims that individuals had natural rights that could be viewed as inherent in the individual. Legal Positivists, and others, claimed that natural rights lacked a scientifically discernable foundation. While the law of gravity could be empirically observed by anyone anywhere and at anytime, the idea that a Creator had created people with certain inalienable rights was rejected as unscientific. The outcome of this development in Germany by the 1930s was that justice was reduced to legal procedures. Substantive justice was dismissed as unscientific. Consequently when Adolph Hitler came to power the laws passed by the Nazi Party to restrict the rights of Jews and many others were viewed as procedurally just. The utter lack of substantive justice in Hitler's "Final Solution of the Jewish Problem" and the slaughter of millions of Jews and others presented the post-war era with a serious challenge.

The challenge was met with a new terminology—human rights. Internationally human rights are the civil rights and civil liberties that are possessed by every person even if these rights have to be asserted as an article of faith.

Growth of Human Rights As Civil Rights

Since 1945 there has been an enormous growth in rights. A number of international agreements have been adopted by the United Nations or by other organizations that list in

detail the rights of persons. The United Nations Declaration of Human Rights (1949) can be viewed as the beginning of a modern international movement for human rights. Influenced by socialist and communist philosophies the Declaration literally expanded the scope of civil rights to include vast new fields of social and economic rights.

Additional UN treaties and conventions on such topics as genocide, refugees, and the environment have been matched by regional treaties on human rights by groups of countries. To protect these extensions of rights new courts for human rights or special offices for protecting civil rights have been created. For example the European Union has created the European Court of Human Rights based at Strasbourg.

The development of the EU since the mid-1950s has created in Europe and well beyond bundles of civil rights for citizens of EU states. These rights have recognized by treaties and have recognized the rights of migrant workers (and following the *Cassis de Dijon* case, allowed for the free movement of goods--a kind of property right), of non-workers, such as students, retirees, and the independently wealthy. Furthermore, as the idea of a European citizenship has grown it has brought a bundle of new civil rights into life for the protection of people. Some of these are political rights to run for office and vote where ever a European is resident in elections for the European Parliament. In addition citizens of EU member states can be represented abroad diplomatically by other EU member states and their consular or embassy services.

The 20th Century growth in rights has so affected the United States that about one half of the cases decided by the United States Supreme Court in the 20th Century involved civil rights and civil liberties disputes. The American Civil Rights Movement, Feminist Movement and Disabilities Movement have been imitated by other groups seeking a new

form of equality and freedom from discriminatory public policies or by private practices. Civil rights cases do not seem likely to decline in the foreseeable future. This is especially so because, the 20th Century has seen many civil rights advocacy groups created. Undoubtedly these shall find new areas in which to press for greater freedoms in the early decades of the 21st Century.

Conflicting Civil Liberties and Rights

It is not unusual for civil rights and civil liberties to be in conflict. For example, sales people who make phone calls to private residences at dinner time will justify their sales call as a exercise the civil liberty of freedom of speech. However, the person who is disturbed by an unsolicited call will view it as a violation of their civil right to privacy.

It is important to note that both civil rights and civil liberties are really ideological claims to power that have attendant winners and losers. For example civil rights advocates have sought to protect the weak in society from what they perceive as abuse. In the case of children an international convention on the rights of children or state domestic laws that weaken parental control make parents into losers, and it is assumed children winners. Or, extending rights to the handicapped at public expense or at the expense of private organizations operating publicly means that the access to public accommodations is paid for by others. Or the elimination of racial or gender discrimination in places of public accommodation by private persons or private organizations is an expansion of the civil liberties of some at the expense of others.

Free Speech

Freedom of speech is a fundamental civil liberty. Some would call it the most important right of citizens or subjects in a political system. Others would put the right to vote and thus the power to use the ballot to change the

government as paramount. However, it may be decided whether voting or freedom of speech is most important, it is the case that freedom of speech is the freedom to criticize the government. It is hard to imagine a government punishing or suppressing its people for saying wonderful positive things about it or its officials. But, the practice of silencing critics has been and is still wide spread. Governments have long sought to censor opponents.

Freedom of speech is oral, but in practice freedom of the press is simply speech printed or published by some media. At the present time governments which allow nearly unlimited criticism still have limits that are associated with violence. Speech that threatens persons or property is not a protected civil right anywhere.

However, the issue of hate speech has produce attempts to silence hostile opponents. Demanding that speech that expresses moral outrage or angry opposition be suppressed does benefit some at the expense of others. For example Roman Catholic priests in Ireland have been warned recently that it would be an act of hate speech for them to distribute a Vatican document describing homosexual behavior as immoral, unacceptable, and a behavior to be publicly opposed. Or, Protestant ministers in Sweden threatened with prison for reading texts from the Bible which oppose as unrighteous homosexual practice. Examples of "political correctness", of "thought police", of persecuting people whose opinions are unpopular are all too familiar and are unlikely to cease.

In totalitarian regimes freedom of speech as a right is qualified by the nature of the ruling ideology. In recent decades many totalitarian governments have opposed human rights conventions or adopting fuller programs of civil rights and civil liberties for their subjects. They have been joined in their denial that human rights exist by Islamic countries. They

have at times based their opposition on the claim that the idea of rights is a Western cultural imperialism.

Freedom of speech like other civil rights and civil liberties creates areas of freedom from government regulation. Politically this is a choice of values. In the United States and elsewhere the choice has been to allow social control to be reduced in favor of personal freedom, but at the expense of the traditional function of using laws (and other mechanism of social control) to make people virtuous. One notorious area of conflict is government of pornographic or terrorist materials, such as bomb making manuals. The use of censorship is, in the case of pornography, a matter of establishing by official decision what is virtuous and what is not. In the case of political materials it may be a case of public safety, but then it may be a pretext for silencing critics of the government.

Even in the case of regulating the viewing of pornography there is a clash of values between those who claim the right to make such choices and those who reject such claims in the name of the public good. These conflicting claims have become increasingly acute with the growth of the internet. The vast world-wide communications now available allows some to openly advocate claims that a right to total sexual freedom even with children.

Censorship of pornographic materials in all forms of media that involves the use of children is now widely accepted, but challenged. In addition the protection of children from pornographic materials until they reach some age of maturity is also widely enforced, but challenged by many civil libertarians. The ability of computer servers to be based anywhere in the world and yet to be reached anywhere has tempted many governments to engage in censorship. Civil libertarians complain, and perhaps rightly, that pornography is a pretext for developing a program of "regulation" that is actually

censorship whether *de jure* or *de facto* through intimidation.

However, patrons of libraries, and especially those with children, are often outraged by the blatant use of public computers for viewing pornography in a public place. The moral claim in the idea of obscenity (e.g., that which is too filthy for public viewing) is clearly in conflict with the freedom claim of a right to personal hedonism. The past has seen conflicts on this issue in various forms and it is difficult to believe that it is an issue that will not continue for millennia to come.

The enormous growth of computers and the globalization of information and communication through electronic mail or web sites is creating a major civil rights issue over the rights of privacy. For workers in companies many have been shocked to discover that their communications have been examined by their superiors without their knowledge. In examining the issue the courts have usually sided with the owners on the grounds that the one who owns the computer has the right of inspection.

The War on Terrorism has also generated many new civil rights issues. These include the right of detention of combatants, their trial, punishment, and possible execution before military tribunals. In addition shutting down funding systems for terrorism has challenged banking privacy rights. Also the use of warrantless wire taps has raised the ire of civil libertarians. The legislation in the United States, the *Patriot Act* and the *Homeland Security Act* run into the hundreds of pages. These laws describe in technical detail the electronic eaves dropping that is permitted. These laws have been passionately denounced by civil libertarians. Advocates argue that law enforcement is faced with a new situation.

Traditionally law enforcement investigates crimes, but terrorist acts are crimes presenting the need to prevent rather than investigate a devastating event. Prevention, advocates claim,

justifies doing such things as roaming the internet for web sites or for electronic mail that are really in the public domain for signals that are suspicious. Or, to be able to listen to cellular phone conversations without getting a warrant because by the time the warrant has been obtained the suspect will have discarded the cellular phone and obtained another with a new name and number. This in effect assigns the wiretap to the person rather than to the communications object.

Discrimination and Civil Rights

Among the many social conflicts that troubled the Twentieth Century were those involving racial, religious, gender or other forms of discrimination. Discrimination whether practiced as a form of private practice or codified by tradition or by law has existed in numerous forms around the world.

Virtually every society in the world has a top and a bottom. Those at the top, the haves, the elites, royalty, or by whatever standard they are known are in a privileged position compared to those at the bottom. Discrimination per se is not inherently wrong. For example is not likely to be lawful for those with such significant sight impairment that they cannot not see five feet ahead clearly or for ten year olds to be allowed to drive automobiles on public highways. These restriction are reasonable because they can easily be justified by a universal rational standard. In contrast many forms of discrimination cannot be so justified.

Racial, religious, class, gender or other forms of discrimination have been practiced for millennia; however, the future of the continuance of such practices has been challenged by advocates of an equality that demands justification for discrimination. The traditional justification for these discriminatory practices have been falling before an onslaught of civil rights demands from many marginalized groups whether women in the highest circles of society or people long

marginalized because of race, historical events such as the lost of a war, or some other reason.

The enfranchisement of women in the twentieth century is continuing in the twenty-first as a feminist movement that seeks equal rights for women. These demands for full citizenship have even affected many Moslem countries or other traditional societies where women were kept sequestered.

The enormous process of de-colonialization that occurred in the decades following World War II ended imperialism. However, in many cases it set the stage for ethnic or social conflicts between dominant groups and other groups that believed themselves to be victims of discrimination.

Marginalized groups seeking the end to discrimination and full civil rights have include the native peoples of the Americas. In Canada legislation to restore rights to indigenous peoples has been viewed as restorative justice or as an end to discrimination and an expansion of civil rights.

In the United States a variety of act have been passed to redress the losses experienced by Japanese-Americans who were imprisoned during World War II. At times their detention centers were located near Indian tribal lands. The expansion of civil rights for natives peoples has usually been in the form of finally fulfilling promises made in old treaties.

In Mexico the sizeable population of full blooded native people such as Aztecs or in the southern Mexican state of Chiapas have responded to the failure of the government to end discriminatory action which has often included simply ignoring them, by engaging in an armed struggle as well as a well orchestrated political campaign.

Similar movement by indigenous people have sprung up in Central America, and in the Andean regions of South America. The Caribbean countries have long had movements to end discrimination. Marcus Garvey was one of the earliest advocates of equal rights for all

including members of the African Diaspora living in the islands. In some cases the discrimination may be more imagined than real. The effects of poverty usually keep people from advancing. In countries with limited resources few opportunities may be available for improving life.

Around the world whether experienced by the small ethnic groups in the Philippines or among the minor ethnic groups in Africa the effects of discrimination are usually sufficient to deny people their civil rights because those stronger haves seek to resist changes that are threatening to the status quo.

In India the ancient practice of caste discrimination has been officially outlawed since Independence. The caste system worked against menial workers such as those in the Sudra caste. They were victimized by caste discrimination. However, it has been the untouchables who were traditionally considered by upper caste Hindus to be so polluted that they were outside of the caste system that have received the most attention.

The plight of the untouchable was denounced by Gandhi and by others. The post-Independence law that put them on the list of Scheduled Classes sought to end discrimination. The changes have not been well received by some practitioners of a revised form of nationalistic Hinduism. To improve their lot quotas were instituted and other forms of affirmative action were instituted.

Affirmative Action as the Restoration of Civil Rights

Affirmative action programs have expanded enormously to end the effects of discrimination, so that people will not longer be victims of the denial of the civil right to equality of treatment in society. The affirmative action programs have focused on educational access, employment opportunities,

and even on the opportunity to participate in ecological decisions.

In the United States programs to end the effects of past discrimination that denied people equal access to social or economic benefits are called affirmative action programs. Affirmative action programs are called by different names in other countries. In Great Britain and in India they are called “positive discrimination.” In Sri Lanka they are called “standardization. In Nigeria affirmative action programs are called “reflection of the character of the nation”. In “sons of the soil” preferences in Malaysia and in Indonesia.

In many countries affirmative action programs have been created that give preference to preferred groups. Other programs are quota programs that ration social, educational, economic or other benefits by means of some form of selective criteria. Among the countries that have such programs are Australia, Brazil, Canada, Fiji, Israel, India, and Pakistan.

Affirmative action programs are supposed to be programs that create remedial or restorative justice. They often have negative unintended consequences.

For example programs in places such as Kenya that were intended to weaken the influence of Indian merchants have in many cases created Africa owned business that are actually fronts for Asians. In some countries these are called “Ali-Baba” enterprises. In theory Ali is an indigenous person, but Baba is actually the real money behind the enterprise without which it would quickly disappear.

In other cases claimants may allege that they are descended from people who are now in a preferred status. In the United States laws allowing casino or other gambling operations on American Indian reservations have brought a number of people forward claiming that they are Native American descendants who also deserve a reservation—at least large enough to house gambling operations.

Another negative consequence of affirmative action is that such claims for “justice” are often presented as temporary; but observers often conclude that they are like temporary taxes—they never go away. It is to be expected that such programs will continue for some decades and that the effort to abolish these forms of reverse discrimination in the name of justice will encourage significant political battles in future decades.

Governmentally Created Civil Rights

Governments can create civil rights as benefits, but it can also create them as a way to expand participation in the policy making process. Because many environmentally hazardous projects such as land fills for garbage, toxic waste dumps, industrial sites or other similar projects are usually placed on the cheapest suitable lands these locations are often in areas where there are people who are poor. The more affluent people and businesses will usually live and do business in places that are more costly and which do not have toxic threats associated with them. In contrast the poor get whatever they can afford.

In 1991 the People of Color Environmental Leadership Summit adopted “Principles of Environmental Justice.” The “Principles” covered many topics but added racism to its definition. Others have added environmental socioeconomic status, classism, environmental racism, environmental and environmental equity. The idea of environmental justice means the empowerment of poor people in the policy process so that they can have a voice in environmental decisions whether purely public or quasi-public such as the location of pipeline or power lines, or those that involve private corporate developments.

To meet the demands for environmental justice advocates were successful in the Administration of President Bill Clinton in gaining from the Environmental Protection

Agency (EPA) a new policy goal of “environmental justice.” The policy goal was to see that the environment was protected in such a way that the poor or marginalized were included in the environmental decision making process. In the EPA created the Office of Environmental Justice in 1992 to implement the policy. The rather revolutionary form of environmental democracy may well spread in the Twentieth Century.

Futures of Civil Rights and Civil Liberties

The early years of the 21st Century have seen many competing groups struggling over the manner in which people shall be treated by others or by governments. Among governments with a long history of open civil rights and liberties a long struggle is continuing over who may do or say what. The issues of political correctness, hate speech, opposition or advocacy for open homosexual practices, and many others are already in play and the forces engaged do not seem likely to withdraw from their positions. Other issues will involve the rights of persons to move anywhere to work, or to not be economic slaves.

Issues that can be expected to be “in the news” as well as in the courts in many place are sexual issues. These include rights homosexuals to marry and to adopt children; for transsexuals to be included in society on the same conditions as normal people, or for the right to view pornography, or the rights claimed by pedophiles, or for the ending of monogamy, or other similar claims. Besides the assertion of claims to sexual rights are those involving suicide and euthanasia. These claims are part of the claims to privacy or to simple be left alone on the grounds that these and other issues are victimless or simply a matter of personal choice.

In addition the War on Terrorism has revitalized many civil rights and civil libertarian groups. These are seeking to block

governmental efforts to supervise the internet or engage in more invasive government investigations as the governments seek to prevent future terrorist attacks. Or, in some cases, to extend ordinary civil rights to those accused of terrorism. Terrorism presents a new challenge. From a civilian point of view it is a policing issue, while from a national security point of view it is a military issue. In reality it is a combination of both of these that is leading the world, along with civil liberties, into uncharted waters.

Selected References

- Barron, Jerome A. and C. Thomas Dienes. (2000) *First Amendment Law in a Nutshell*. Second Edition. St. Paul, Minnesota: West Group.
- Brysk, Alison. (2002) (Editor) *Globalization and Human Rights*. Los Angeles: University of California Press.
- Chang, Nancy. (2002) *Silencing Political Dissent: How Post-September 11 Anti-Terrorism Measures Threaten Our Civil Liberties*. New York: Seven Stories Press.
- Domino, John. (2003) *Civil Rights and Liberties in the 21st Century*. Second Edition. NY: Addison Wesley Longman.
- Donnelly, Jack. (2002) *Universal Human Rights in Theory and Practice*. Second Edition. Ithaca, New York: Cornell University Press.
- Evans, Tony. (2000) “Citizenship and Human Rights in the Age of Globalization”, *Alternatives*, Volume 25, 429ff.
- Godwin, Mike. (2003) *Cyber Rights: Defending Free Speech in the Digital Age*. Boston: MIT Press.
- Kersch, Ken I. (2003) *Freedom of Speech: Rights and Liberties Under the Law*. Santa Barbara, CA: ABC-CLIO.
- Leone, Richard C. and Gregory Anrig. (2003) (Editors) *The War on Our Freedoms: Civil Liberties in the Age of Terrorism*. New York: Public Affairs.

Pring, George W. and Penelope Canan.(1996) *SLAPS: Getting Sued for Speaking Out*. Philadelphia: Temple University Press.

Randall, Richard S. (2003) *American Constitutional Development: The Rights of Persons*. Volume II. New York: Addison Wesley Longman, Inc.

Sowell, Thomas. (2004) *Affirmative Action Around the World: An Empirical Study*. New Haven, CT: Yale University Press.

Shapiro, Joseph. P. (1994) *No Pity: People with Disabilities Forging a New Civil Rights Movement*. New York: Times Books.

Thierer, Adam and Clyde Wayne Crews. (Editors) *Who Rules the Net?: Internet Governance and Jurisdiction*. Washington, D.C.: Cato Institute.

Vieira, Norman. (1998) *Constitutional Civil Rights in a Nutshell*. Third Edition. St. Paul, Minnesota: West Group.

Weinstein, James. (1999) *Hate Speech, Pornography, and the Radical Attack on Free Speech Doctrine*. Bolder CO: Westview Press.

Walter, Samuel. (1999) *In Defense of American Liberties: A History of the ACLU*. Second Edition. Carbondale, Illinois: Southern Illinois University Press.

Zelezny, John D. (2003) *Communications Law: Liberties, Restraints, and the Modern Media*. Belmont, California: Wadsworth.

Websites

American Civil Liberties Union. www.aclu.org

Americans for Democratic Action. www.adaction.org

Disability Rights Education and Defense Fund. www.dredf.org

Guide to Human Rights. www.hg.org/human.html

Human and Civil Rights: Internet Resources. internet.ggu.edu/university_library/humanrights.html

Human Rights Organizations.

hotburrito.100megsfree5.com/links/rights.html

Human Rights Watch. www.hrw.org

Privacy International. www.privacyinternational.org

UN Agreements on Human Rights. www.hrweb.org/legal/undocs.html.

University of Minnesota Human Rights Library: www1.umn.edu/humanrts/bibliog/BIBLIO.htm

United States Civil Rights Code. www4.law.cornell.edu/uscode/42/ch21.html.

United States Commission on Civil Rights. www.usccr.gov

United States Department of Commerce. Office of Civil Rights: www.osec.doc.gov/ocr

Yahoo Directory of Human Rights Organizations: dir.yahoo.com/society_and_culture/issues_and_causes/human_rights/organizations/

Andrew J. Waskey
 Department of Social Sciences
 Dalton State College
 Dalton, Georgia, USA
jwaskey@daltonstate.edu

Corporate Social Responsibility

Eva E. Tsahuridu

Introduction

There is limited agreement about the content and purpose of corporate social responsibility (CSR). Some people perceive it as concentrating on philanthropy, doing good to appease or assist society. Others see it as an ethics and values approach to business which requires the redefinition of the purpose and processes of business corporations. Corporate social responsibility and its related concepts and terminology such as accountability, sustainability, citizenship, responsiveness and stakeholder management are very recent phenomena, which developed in the second half of the twentieth century. The increase in CSR interest and activity is attributed to many factors, such as the increased power of corporations and decreased power of governments which makes corporations a lot more influential, the increased shareholder and other stakeholder activism which demands more responsible corporate behaviour, the increased legal and regulatory developments which prescribe more socially responsible behaviour, and increased concerns for the environment and the impact of corporations on it.

In addition, there is increased questioning of the economic assumption of self-interest as the supreme motivation of human behaviour, which has permeated the management of corporations and markets, and an increased realisation that moral and social concerns are just as important as economic concerns. CSR and its related concepts pertain to the external legitimacy of the corporation because they address the impact corporations have on society; while corporate governance relates to its internal legitimacy, concerned with questions such as who runs the corporation,

for whom and by what means (Epstein, 1999).

Carroll (1994), one of the important contributors to the CSR field, describes it as a field with a poor map, with loose boundaries; a multidisciplinary field without focus, which consists of people from diverse backgrounds and with different perspectives. A synthesised definition of CSR from many of its current advocates such as Donaldson, Selznick, Evan and Freeman, Paine, Ulrich, Carrol, is provided by Dubbink (2004), who describes it as the opinion that market organisations have a responsibility for public issues. This responsibility extends further than the limits set by law and common decency.

It is not clear whether CSR is a process or an outcome and where its boundaries lie. Originally CSR was perceived as an outcome, which was to make the corporation socially responsible. Later it developed into a process with Jones (1980) being the first to argue that CSR includes CSR decision making and CSR behaviour. CSR as a process creates more considerations that need to be addressed by corporations relevant, such as awareness, language, values and expectations.

The lack of agreement about what corporate social responsibility is, leads to a number of views and approaches. The two extreme views are that corporations have no other responsibility beyond profit maximisation within the legal framework and organisations have additional responsibilities relating to society in addition to making profits. The voices that are aligned with the latter view, the social responsibility view, are growing in number and influence. This growth may represent a general increase in interest, awareness and changed expectations for the behaviour of corporations.

Why Do Business Organisations Exist?

Milton Friedman's (1970) thesis that "the social responsibility of business is to increase

its profits” epitomises the view of the amorality of business and the primacy of the shareholder. Friedman argues that using shareholders’ money for anything other than making profits for them is wrong. He supports the neoclassical economists’ amoral thesis, which is based on egoism. This view is attributed to the premise that the common good is best achieved by the individual pursuit of self-interest and business profits, rather than by actions based on conscious moral purpose (Steiner & Steiner 1991).

More recently the ethics or value-based approach to CSR developed. Morality is based on the premise that a person utilises ego capacities for ethical rather than egoistic ends (Hoffman cited in Shelton, & McAdams 1990). In business, there seems to be a different understanding of egoism, and the requirement for self-interest makes egoistic ends desirable and necessary and self-interest good and valuable. The current political economy and business is based on that premise as is Friedman’s position. Friedman’s position is based on the fallacy that shareholders are motivated entirely by self-interest and seek to maximise profits but managers are not (Grant 1991). Managers, the agents of the self-interested shareholders, are assumed to be totally devoid of self-interest or their self-interest is expected to be contained so they can be dedicated to the self-interest of the shareholders.

Recent organisational failures that resulted from management misbehaviour (Enron, WorldCom, Tyco, Parmalat are some indicative examples) confirm the paradoxical nature of Friedman’s thesis. What is rediscovered is that managers too can be guided by self-interest and can use any means to achieve their selfish ends. What they are in fact doing is applying the methods and processes that were acceptable for use to increase shareholder returns to themselves, to further their own financial and other interests.

Another issue that is raised from Friedman’s position and the general CSR dialogue is for whose benefit the business corporation exists. The maximisation of the return to the shareholder as the main objective of the corporation is increasingly questioned. Handy (2002) calls the fact that shareholder needs are confused with the purpose of the corporation a logical confusion. He claims that shareholders are not owners but they are investors or even gamblers, and to turn their needs into the corporate purpose is a mistake. Similarly Phillips, Freeman, and Wicks (2003) comment that to perceive shareholders as identical to the corporation, is wrong. The corporation is a distinct entity and its management is not the agent of shareholders, but rather management is the agent of the organisation. Management, as a result, needs to behave in ways that improve the long term viability of the organisation.

Duska (1997) emphasises the difference between the motives of corporations and their purpose, recognising the motive of corporations to be profit but their purpose to be the provision of goods or services. Duska (1997:197) argues that the view that the sole responsibility of a business is profit maximisation is “an insidious mistake”. Society accepts business organisations because they provide benefits, but “no society would permit a system that did it more harm than good. The appeal to profit was a means to motivate more production but it was not the purpose of the production” (Duska 1997:198). Similar sentiments and beliefs on the value of economic efficiency were expressed by Tawney (1926:277) early this century: ‘Economic efficiency is a necessary element in the life of any sane and vigorous society...but to convert efficiency from an instrumental into a primary object is to destroy efficiency itself’.

The capacity of profit to measure performance is also an issue that is increasingly addressed (Valor 2005) because profit does not measure positive and negative externalities. Economic profit is, as a result, an inefficient measure of corporate performance.

Wheeler, Colbert and Freeman (2003) argue that the primary motivation of corporations is the creation of value. They accept the different understandings of what value is from the different actors that are involved in the economy. They offer three questions for business propositions to test whether they are indeed valuable. They argue that a value proposition needs to be feasible, it needs to be supported by its stakeholders and it needs to be economically, environmentally and socially sustainable in the long term.

Moral Personhood of Corporations

The prominence of and interest in CSR and ethics led to an increased interest in the ontology of organisations in general and business corporations in particular. The ontology of corporations shapes their responsibilities. The ongoing debate of what corporations are, relates to the moral personhood of corporations. Moral personhood contains moral agency, and moral agency contains moral autonomy and responsibility. If corporations are moral agents then they also have moral responsibilities and they can be evaluated in moral terms as well as economic terms.

The issue of CSR is based on different views of the corporation. Generally in the literature there are attempts to understand whether the corporation is a moral entity or merely a structure that provides the context where moral entities, the people of corporations, behave. This debate enlightens the questions of how can a corporation have

responsibilities and what kind of responsibilities it can have.

The three main views of the ontology of corporations perceive it as moral persons, as property and as partial moral persons. In addition to these views, corporations are also perceived as communities or moral worlds. This last view does not perceive corporations as persons, but accepts the influence they have on people and groups that decide and act in and for them.

The moral person view of the corporation is supported by Clinard and Yeager (1980), French (1979, 1996), Weaver (1998), and Sandelands and Stablein (1987) among others. This view attributes moral personhood to corporations. The existence of an internal decision making structure with policies, rules and procedures in corporations, their capability to perform intentional actions, in and of themselves, their ability to use language and adaptability to multiple personalities, are characteristics used to support their moral agency. These characteristics make possible the subordination of individual action to corporate action.

Sandelands and Stablein (1987) further claim that corporations are mental entities capable of thought. They claim that the premise adopted by many organisational theorists that corporations do not make decisions only people do, limits them to only examine decisions in corporations without ever considering the possibility of decision making by corporations. In contrast, the organisation mind concept suggests that to understand decision making in corporations, it is not enough to describe what is in the minds of the members of the corporations, as individuals may know more and less than organisations (see Weick & Roberts 1993).

The antithesis to the view that corporations are moral persons is the structural restraint view, the view that

corporations are nothing more than property. This view perceives corporations as artificial persons and as such possessing only artificial responsibilities. “[B]usiness’ as a whole cannot be said to have responsibilities” claims Friedman (1970:126), only persons can have responsibilities. Ladd also supports this view, while Ewin (1991) sees the moral personality of corporations as severely limited and exhausted by their legal personality. The personality of corporations for Ewin is restricted to requirements, rights and duties, and not one that is capable of virtue and vice.

Ladd (1970) supports the structural restraint position and claims that the principle of the exclusion of the irrelevant is part of the language game and reveals that morality is not relevant in organisational behaviour. The language game of social decisions permitted actions to be attributed to corporations rather than individuals, but it did not contain concepts like “‘moral obligation’, ‘moral responsibility’, or ‘moral integrity’” according to Ladd (1970:119). These terms are, however, found in the contemporary lexicon of corporations. Ladd (1970) differentiates between corporate acts and personal acts based on the goal they are directed towards. Ladd, claims Heckman (1992), determines good and bad actions by the achievement of organisational goals. He thus considers any act that does not lead to goal attainment an individual act and any act that leads to the goal attainment of corporations, a good act. This is a result of seeing corporations as ends in themselves and not as entities that exist to benefit society, a consequence of the separation thesis (Freeman 1994). The separation thesis perceives business and ethics as independent and categorically distinct realms, with different concepts, language and logic.

These phenomena lead to the perception that corporate actions are to be assessed solely on goal accomplishment and not on

moral responsibility or moral behaviour, thus eradicating the possibility of a bad organisational act. Ladd concedes, however, that the moral schizophrenia of corporate ‘rationality’ and individual morality must be resolved by somehow surrendering neither. Unlike the moral person view, Ladd claims that a corporation is unable to consider moral issues in its decision making, thus making it more similar to a machine rather than a moral agent. Ladd’s view appears congruent with the amoral calculator model of decision making described by Vaughan (1998:26). She explains that the amoral calculator is evident ‘when an organisation experiences structural strain to achieve its goals, individuals acting in their organization roles weigh the costs and benefits of their actions, choosing to violate laws and rules to attain organizational goals’.

The third view on the moral personhood of corporations sees them as partial moral persons and attributes secondary moral agency to them. It holds both corporations and persons responsible for actions. Nagel (1979), for example, argues that the guilt for public wrongdoings, which is what corporate actions are, may be attributed to individuals just as private wrongs. The responsibility of the public wrong, however, is partly absorbed by the moral defects of the corporation through which the act is undertaken. The responsibility that can be attributed to the corporation, he claims, is in inverse relationship to the power and independence of the actor. Another view is that corporations possess restricted personhood (Nesteruk & Risser 1993). They possess personhood because the corporation is a moral agent due mainly to its internal decision making structure, but it can also be understood as property in the service of human interests. Corporations are not, however, accepted as property in the service of all humans, and in many instances the interests of the individuals within the corporation who make the

decisions are not of concern, because as agency theory claims corporations are managed to satisfy the needs of the principals, the shareholders.

Werhane (1989) views corporations as collective secondary moral agents because although they cannot act, they create anonymous policies and practices that are not traceable to individuals, but upon which corporate activities are based. Velasquez (1992) also sees corporations as having moral duties and moral responsibilities in a secondary sense. Similarly, Wilmot (2001) sees corporations as having moral agency and as such moral responsibility but a responsibility that is limited because it depends on a more limited autonomy. The partial moral person view perceives individuals who underlie the corporation as the primary bearers of moral duties and responsibilities and reduces corporate activity to individual contribution.

“Human individuals are responsible for what the corporation does because corporate actions flow wholly out of their choices and behaviours” (Velasquez 1992:19). De George (1990) restricts the moral duties and responsibilities of corporations to the avoidance of immoral ends for which they are formed and immoral means by which these ends are pursued. Corporations, he claims, can not be expected to act from moral motives but are expected to avoid behaving immorally. This allows corporations to be amenable to moral evaluation in the absence of moral personhood per se.

Frederick and Weber (1987) attribute moral responsibility for corporate acts to corporations and individuals. Personal values, according to them, are involved but may not be central to decisions and actions, because they only constitute a portion of the total value structure of a corporation. The corporation is thus morally responsible for corporate acts because it has its own values

and traditions, and not the individuals who make and carry out decisions in corporations. This does not extinguish individuals’ responsibility because they agree to abide with the corporation’s rules and procedures.

Metzger and Dalton (1996), after reviewing the debate of organisational moral agency, conclude that those who deny corporations moral agency on the grounds that they insufficiently resemble human beings, need to subject their assumptions about human beings to more rigorous scrutiny.

Corporate moral personhood is an important issue because if we accept the moral personhood of corporations, then we must hold them solely accountable and responsible for their actions. This will limit individual responsibility for ethical misconduct in corporations. If, however, we accept corporations as structures only, then we do not accept the corporation as a being, and see it only as a structure in which people decide and act. In this case moral responsibility is attributed solely to the persons in the corporation. Recent writing and theorising is more likely to attribute some moral responsibility to the corporation.

The main reason for the denial of corporate moral agency is the fear of diluting personal moral responsibility (Metzger & Dalton 1996; Werhane 1989). Werhane (1989), however, argues that corporate moral responsibility does neither limit nor reassign personal moral responsibility but extends it to the corporation and its policies and practices.

CSR is attributable to the corporation not the individual in the corporation, so the unit of analysis is the corporation. As a consequence, it remains silent as to the responsibilities of the individual in the corporation, but holds corporations accountable for their behaviour. Thus, CSR attributes some moral agency to the corporation.

Theories of Corporate Social Responsibility

Four groups of CSR theories are identified by Garriga and Mele (2004). These are the instrumental theories, political theories, integrative theories and ethical theories.

The *instrumental theories* perceive CSR as a means of achieving economic objectives and the ultimate goal of wealth creation. This is where the view of the maximisation of shareholder wealth belongs and Friedman's (1970) thesis finds a home. In the instrumental group of CSR theories there are also those that aim to achieve competitive advantage for the corporation and long term social objectives; and those that are concerned with cause-related marketing and aim to differentiate a product by creating socially responsible attributes. These attributes affect the corporation's reputation, as people are more likely to perceive it as reliable and honest and its products as having higher quality.

The *political theories* of corporate constitutionalism and integrative social contract theory, focus on the use of business power in the political sphere. Corporate constitutionalism is concerned with the power corporations have in society and the responsibilities they have in using their power. The integrative social contract theory looks at the relationship between the corporation and society as a social contract (Donaldson & Dunfee 1999). In this theory social responsibility results from consent at the macrosocial and microsocial levels. Macrosocial contracts consist of rules that apply to all contracts and provide the hypernorms. Hypernorms are "fundamental moral precepts for all human beings ... which express principles so fundamental to human existence that one would expect them to be reflected in a convergence of religious, political, and philosophical thought"

(Donaldson & Dunfee 1995:95–96). Microsocial contracts show agreements that are binding in a community and in order to be legitimate they need to agree with the hypernorms. Political theories show that corporate citizenship changed from describing the relationship between the corporation and society to the description of the responsibilities of corporations towards the local community and the environment. It is argued that this development in citizenship is a consequence of government's failure to protect citizens, and the increased power of corporations.

The *integrative theories* focus on the integration of social demands, which form the interaction between society and the corporation. Issues management, the principle of public responsibility, stakeholder management and corporate social performance are theories that constitute the integrative theories.

Carroll's (1991) contribution to CSR is part of the integrative approach as it seeks to legitimise corporate existence. Carroll conceptualises CSR as a pyramid which contains four kinds of responsibilities that corporations need to consider. These responsibilities are economic, legal, ethical and philanthropic. Philanthropic responsibilities are the only ones that are discretionary or voluntary. Carroll and Buchholtz (2003) describe the economic and legal responsibilities as required, the ethical as expected and the philanthropic as desired. The recent legal and regulatory developments, however, seem to make ethical responsibilities required and not only expected. Carroll's view of CSR prescribes ethical responsibilities to the corporation and extends their obligations beyond legal profit maximisation. Ethical responsibilities include doing what is right and also doing what is good, that is being ethical in terms of means and ends, processes and outcomes.

The *ethical theories* of CSR focus on business doing what is right in order to do good for society. The normative stakeholder theory, universal rights, sustainable development and the common good are approaches that focus on corporations doing what is ethical.

CSR places certain demands on corporations. Dubbink (2004) describes four such demands that are developed in the CSR literature. The first demand is to comply with the law and common decency. Corporations must also comply with the other demands in order to fulfil their CSR obligations. The second demand is to ensure the adequacy of laws. The third demand of CSR is to inflict no harm. The fourth demand is to be sympathetic towards the issues society raises that relate to the actions of the corporation. These demands reflect an alteration in the expectations and demands from corporations and they show that maximising wealth is no longer considered an adequate contribution to society and a fulfilment of the corporations' responsibilities.

Stakeholder Theory

Stakeholder theory is an integrative theory of CSR that is gaining in popularity and is found extensively in the corporate, but also in the political and social lexica. It is one of the three leading normative theories of business ethics, the other two being the shareholder and social contract theories (Hasnas 1998). Hasnas describes the stakeholder theory as both empirical and normative. As an empirical theory, stakeholder theory prescribes a method for improving the performance of business, whilst as a normative theory it asserts that regardless of the effect on business performance, the business organisation should benefit all its stakeholders. Garriga and Mele (2004) describe stakeholder theory as both integrative and ethical. Integrative because it

aims to integrate demands made by groups that have a stake in the organisation and ethical because it has a normative ethical core.

The main prescription of stakeholder theory is that corporations ought to be operated for the benefit of all those who have a stake in them. It provides an antithetical premise to the shareholder view of the firm which claims that corporations ought to be operated for the benefit of their shareholders only.

The aim of stakeholder theory is to address managing and ethics. It aims to overcome the amorality (the absence of ethics) of business that resulted from the separation of ethics from economics and also address the immorality of business. Stakeholder theory has explicit moral content and looks at the means and ends of corporate activity (Phillips et al 2003). Stakeholders are not only perceived in an instrumental manner because of the effect they can have on the business organisation, but they are also examined in a normative sense.

Donaldson and Preston (1995) describe stakeholder theory as descriptive, describing corporations as constellations of interconnected interest groups; instrumental, suggesting that corporations that adopt stakeholder management will be more successful than those that do not (other things being equal); managerial, by enabling managers to identify options and solutions; and normative, by developing moral and philosophical guidelines for the operation of corporations.

Stakeholder theory distinguishes between different classes of stakeholders. Generally they are called primary and secondary or normative and derivative (Phillips 2003). Primary or normative stakeholders are the stakeholders for whose well-being the organisation has a direct moral obligation. Employees, financiers, customers, suppliers

and the local community are generally considered normative or primary stakeholders. Derivative or secondary stakeholders have the capacity to harm the organisation or benefit from it, but the organisation is not managed explicitly for their benefit. Competitors, activists and the media are placed in this category. Even though the organisation is not managed for the direct benefit of derivative stakeholders, they must be taken into account if they affect the organisation or its normative stakeholders.

Stakeholder theory is concerned with fairness in the effect the organisation has on all the stakeholders. This fairness is assessed not only in terms of outcomes but also processes; thus stakeholder theory is concerned with distributive and procedural justice. The fairness in process is achieved through participation and involvement, while the fairness in outcomes is achieved by looking at the contributions of each stakeholder group.

Phillips, Freeman, and Wicks, (2003) emphasise the fact that stakeholder theory is a theory of strategy and ethics and not a theory of political economy. As such it does not aim to alter the existing political or economic system. Instead it provides ethical and strategic prescriptions for the existing political and economic system without questioning the soundness of the system.

Concerns about CSR

There are a number of criticisms of CSR, some of them contradictory. CSR is criticised for its narrow focus and for its broad focus, for its academic routes, for the difficulty in its operationalisation, for its attack on property rights and its threat to free society (see Valor 2005).

CSR and its increased prominence is questioning not only the obligation of corporations and the appropriateness of their behaviour, it also questions the Anglo Saxon

political and economic system, it questions capitalism and the free market. Dubbink (2004) develops the incompatibility of CSR with the dominant political theory and the free market system. This incompatibility stems from the collision between the demands of CSR and political theory: the former demanding virtuous behaviour in the market, the latter inhibiting it and demanding self-interested behaviour.

Valor (2005) presents three arguments on CSR. The descriptive approach asks for ethical values to be incorporated in economic decisions, the instrumental approach emphasises profit maximisation and perceives ethical values as a constraint and the third approach, the normative, requires corporations to see social performance as an end in itself. The last approach is the one that requires not only changes within the existing economic system but change in the system itself. Valor argues that system change requires society to change, as the system reflects the values and demands of society. Valor claims that the CSR discourse has been added on the neoclassical view. This addition however did not result in the alteration of the features and the aims of the corporation. As a result, CSR is perceived as another tool to improve the competitiveness and profitability of the firm. This bandage approach does not alter nor challenge the primacy of the shareholder as the only stakeholder who matters in corporate affairs. The outcome of this approach is to increase the instrumental use of CSR without altering the objectives and goals of business organisations. This approach seems to be adopted by Drucker (1984:62) who claims that the proper meaning of CSR is to 'tame the dragon, that is to turn a social problem into economic opportunity and economic benefit, into productive capacity, into human competence, into well-paid jobs, and into wealth'. Another concern with CSR is the fact that it is

primarily contained in the rhetoric of corporations and does not infiltrate their philosophy, processes and objectives. Roberts (2003) calls it a prosthesis, a cheap and easy approach that can easily be attached but in appearance only, not content or substance. These concerns are raised primarily for the instrumental or the 'business case' of CSR.

The instrumental use of CSR leads to what is called in the literature (see Valor 2005) 'managerial capture'. Managerial capture refers to actions of management, on behalf of the corporation, to control and redefine the meaning of CSR and ensure it remains wealth maximising. The use of communication to advance the corporate image through the means of CSR, or alternatively, the unwillingness of management to trade profits for doing good are outcomes of managerial capture.

Measurement of CSR

The increased interest on the social responsibilities of corporations led to a dramatic increase of CSR measurement and ratings. This development has in turn led to an increase of internal specialists that measure and communicate their corporation's performance (Marquez & Fombrun 2005). Marquez and Fombrun explain that the rise in social investment funds and social regulations, further oblige corporations to address the social and environmental impact of their activities and these phenomena further increase the explosion of CSR ratings.

There are numerous criteria that are used to rate a corporation's social performance, such as employment practices, social and environmental impact, governance etc. A number of rating agencies exist such as the SIRI group, EIRIS and the ECGS network. The tension between standardisation and differentiation in CSR ratings is evident.

Triple bottom line accounting is another recent development that seeks to expand the

traditional corporate reporting framework from the economic to the environmental and social performance as well. Numerous codes and regulations such as the OECD guidelines, the Global Reporting Initiative (GRI) and the AA1000, address triple bottom line accounting and an increasing number of companies are adopting it. Concern is raised, however, about the soundness, ability to operationalise and appropriateness of the concept (see Norman & MacDonald 2004).

Conclusion

The move towards the moralisation of business in the postindustrial society has been assisted by the development of the organisational moral personhood and the development of CSR. Corporations are now accepted more than ever before by theorists, consumers and the law as anthropomorphous organisms with imagination, learning, memory, character, knowledge, reputation and identity. They are accepted as members of society with rights and obligations, members with virtuous and vicious capacities. CSR provides a vehicle of achieving corporate moral behaviour.

Traditionally, corporations saw their purpose in terms of responding to the changing preferences and needs of consumers, or some argue in shaping consumer preferences and needs. Recently however, the emphasis has moved to the compliance with the social, ethical and legal values of society. The preferences and needs of consumers remain the principal interest of most corporations, but valuable for consumers is not necessarily only the economic and the material, but increasingly also the socially responsible, the humane, the environmentally sensitive and sensible. Society, increasingly expects the congruency of its values with corporate values, more than it has done since the industrial revolution and it expects corporations to behave in

accordance with these values. Over two decades ago the opinion was expressed that "few issues promise to have more long-run impact on both business and society than those of corporate attitudes toward social responsibility, corporate behaviours in response to such attitudes, and societal replies to those behaviours" (Aldag & Jackson 1977:65). Today, business and society are in dialogue about corporate social responsibility.

Selected References

- Aldag, R.J. and D.W. Jackson Jr. (1977) "Assessment of Attitudes Toward Social Responsibilities", *Journal of Business Administration*, Volume 8, Number 2, pp. 65-80.
- Carroll, A.B. (1991) "The Pyramid of Corporate Social Responsibility: Toward the Moral Management of Organizational Stakeholders", *Business Horizons*, Volume 34, Number 4, pp. 39-48.
- Carroll, A.B. (1994) "Social Issues in Management Research", *Business & Society*, Volume 33, Number 1, pp. 5-25.
- Carroll, A.B. and A.K. Buchholtz. (2003) *Business and Society: Ethics and Stakeholder Management. Fifth Edition*. Cincinnati: South-Western College.
- Clinard, M.B. and P.C. Yeager. (1980) *Corporate Crime*. New York: The Free Press.
- De George, R.T. (1990) *Business Ethics*. Third Edition. New York: Macmillan Publishing Company.
- Donaldson, T. and T.W. Dunfee. (1999) "When Ethics Travel: The Promise and Peril of Global Business Ethics", *California Management Review*, Volume 41, Number 4, pp. 45-63.
- Donaldson, T. and L.E. Preston. (1995) "The Stakeholder Theory of the Corporation: Concepts, Evidence and Implications", *Academy of Management Review*, Volume 22, Number 1, pp. 65-91.
- Drucker, P.F. (1984) "The New Meaning of Corporate Social Responsibility", *California Management Review*, Volume 26, pp. 53-63.
- Dubbink, W. (2004) "The Fragile Structure of Free Market Society: The Radical Implications of Corporate Social Responsibility", *Business Ethics Quarterly*, Volume 14, Number 1, pp. 23-46.
- Duska, R.F. (1997) "The Why's of Business Revisited", *Journal Of Business Ethics*, Volume 16, pp. 1401-1409.
- Epstein, E.M. (1999) "The Continuing Quest for Accountable, Ethical and Humane Corporate Capitalism", *Business & Society*, Volume 38, Number 3, pp. 253-267.
- Ewin, R.E. (1991) "The Moral Status of the Corporation", *Journal Of Business Ethics*, Volume 10, pp. 749-756.
- Frederick, W.C. and J. Weber. (1987) "The Values of Corporate Managers and Their Critics: An Empirical Description and Normative Implications", *Research in Corporate Responsibility and Policy*, Volume 9, pp. 131-152.
- Freeman, R.E. (1994) "The Politics of Stakeholder Theory: Some Future Directions", *Business Ethics Quarterly*, Volume 4, Number 4, pp. 409-421.
- French, P. (1979/1988) "The Corporation as a Moral Person", in T. Donaldson and P.H. Werhane (Editors), *Ethical Issues In Business: A Philosophical Approach*. Third Edition. New Jersey: Prentice Hall, pp. 100-109.
- French, P.A. (1996) "Integrity, Intentions and Corporations", *American Business Law Journal*, Volume 34, Number 2, pp. 141-155.
- Friedman, M. (1970/1984) "The Social Responsibility of Business is to Increase Its Profits", in W.M. Hoffman and J.M. Moore (Editors), *Business Ethics:*

- Readings and Cases In Corporate Morality*. New York: McGraw-Hill, 126-131.
- Garriga, E. and Mele, D. (2004) "Corporate Social Responsibility" *Journal of Business Ethics*, Volume 53, pp. 51-71.
- Grant, C. (1991) "Friedman Fallacies", *Journal of Business Ethics*, Volume 10, pp. 907-914.
- Handy, C. (2002) "What's a Business For?", *Harvard Business Review*, Volume 24, December, pp. 49-54.
- Hasnas, J. (1998) "The Normative Theories of Business Ethics: A Guide for the Perplexed", *Business Ethics Quarterly*, Volume 8, Number 1, pp. 19-42.
- Heckman, P. (1992) "Business and Games", *Journal of Business Ethics*, 11, 933-938.
- Jones, T. M. (1980) "Corporate Social Responsibility Revisited, Redefined", *California Management Review*, Spring, pp. 59-67.
- Ladd, J. (1970/1988) "Morality and the Ideal Of Rationality in formal Organizations", in T. Donaldson and P. H. Werhane (Eds.), *Ethical Issues In Business: a Philosophical Approach*. Third Edition. New Jersey: Prentice-Hall, 110-122.
- Marquez, A. and C.J. Fombrum. (2005) "Measuring Corporate Social Responsibility", *Corporate Reputation Review*, Volume 7, Number 4, pp. 304-308.
- Metzger, M. B. and D.R. Dalton. (1996) "Seeing The Elephant: An Organizational Perspective on Corporate Moral Agency", *American Business Law Journal*, Volume 33, pp. 489-576.
- Nagel, T. (1979) *Mortal Questions*. Cambridge: Cambridge University Press.
- Nesteruk, J. and D.T. Risser. (1993) "Conceptions of the Corporation and Ethical Decision Making in Business", *Business & Professional Ethics Journal*, Volume 12, Number 1, pp. 73-89.
- Norman, W. and C. Macdonald. (2004) "Getting to the Bottom of 'Triple Bottom Line'". *Business Ethics Quarterly*, Volume 14, Number 2, pp. 243-262.
- Phillips, R. (2003) "Stakeholder Legitimacy", *Business Ethics Quarterly*, Volume 13, Number 1, pp. 24-41.
- Phillips, R.; R.E. Freeman and A.C. Wicks. (2003) "What Stakeholder Theory is Not", *Business Ethics Quarterly*, Volume 13, Number 4, pp. 479-502.
- Roberts, J. (2003) "The Manufacture of Corporate Social Responsibility: Constructing Corporate Sensibility", *Organization*, Volume 10, Number 2, pp. 249-265.
- Sandelands, L.E. and R.E. Stablein. (1987) "The Concept of Organization Mind", *Research in the Sociology Of Organizations*, Volume 5, pp. 135-161.
- Shelton, C.M. and D.P. McAdams. (1990) "In Search of an Everyday Morality: The Development of a Measure", *Adolescence*, Volume 25, Number 100, pp. 923-943.
- Steiner, G. and J. Steiner. (1991) *Business, Government and Society: A Managerial Perspective*. Sixth Edition. New York: McGraw Hill.
- Tawney, R. H. (1926) *Religion and the Rise of Capitalism*. Middlesex: Penguin Books.
- Valor, C. (2005) "Corporate Social Responsibility and Corporate Citizenship: Towards Corporate Accountability", *Business and Society Review*, Volume 110, Number 2, pp. 191-212.
- Vaughan, D. (1998) "Rational Choice, Situated Action, and The Social Control Of Organizations", *Law and Society Review*, Volume 32, Number 1, pp. 23-61.
- Velasquez, M.G. (1992) *Business Ethics: Concepts and Cases*. Third Edition. New Jersey: Prentice Hall.
- Weaver, W.G. (1998) Corporations As Intentional Systems. *Journal of Business Ethics*, Volume 17, pp. 87-97.

- Weick, K.E. and K.E. Roberts. (1993) "Collective Mind in Organizations: Heedful Interrelating on Flight Decks", *Administrative Science Quarterly*, Volume 38, pp. 357-381.
- Werhane, P.H. (1989) "Corporate and Individual Moral Responsibility: a Reply To Jan Garrett", *Journal of Business Ethics*, Volume 8, pp. 821-822.
- Wheeler, D.; B. Colbert and R.E. Freeman. (2003) "Focusing on Value: Reconciling Corporate Social Responsibility, Sustainability and a Stakeholder Approach in a Network World", *Journal of General Management*, Volume 28, Number 3, pp. 1-28.
- Wilmot, S. (2001) "Corporate Moral Responsibility: What Can We Infer From Our Understanding of Organisations?", *Journal of Business Ethics*, Volume 30, pp. 161-169.

Eva E. Tsahuridu
Department of Management
University of Greenwich
London, UK
eva.tsahuridu@rmit.edu.au

Counterfeiting

Edward O'Boyle

Introduction

Counterfeiters routinely take pains to avoid detection, and to the extent that they are successful the total losses due to counterfeiting are no more than guesstimates. In 1983, for example, annual losses in the United States according to an investigative committee of the U.S. House of Representatives were estimated at \$20 billion (O'Connell et al 1985:65). The same \$20 billion figure is reported by the International AntiCounterfeiting Coalition (IACC) for 1994 (NAPM 2000:1). The U.S. Department of Justice estimates that in 2002 losses to U.S. businesses due to counterfeiting amount to \$200-250 billion a year (USDJ 2002:1).

The OECD (1998:4) estimated that counterfeiting in 1998 accounted for five percent of world trade due mainly to technological advances that have made counterfeiting easier, increased global trading, and a larger market share of products such as branded clothing and software that are attractive to copy. Citing the International Chamber of Commerce (ICC), another source put the yearly global counterfeit trade at five to seven percent of worldwide trade or \$450-500 billion (O'Brien 2003:1). The ICC's own Commercial Crime Services affirmed the five-seven percent range as correct but reduced the dollar volume of global counterfeit trade to approximately \$400 billion (Lowe 2003).

In a report published in 2008 OECD estimated world trade in counterfeits in 2005 at \$200 billion (OECD 2008:13). On the other hand, IACC put the counterfeit trade at \$600 billion (IACC 2008:1). Amid this confusion and widely different estimates which are fostered by the clandestine nature of counterfeiting, what we can say with

confidence is that counterfeiting has become such a major global problem that in recent years several international organizations including the IACC (1978-79), the ICC's Counterfeiting Intelligence Bureau (1985), and the Pharmaceutical Security Institute (2001) have been established explicitly to deal with it. In 1995 the World Trade Organization implemented the most comprehensive multilateral agreement known as TRIPS (Trade-Related Aspects of Intellectual Property) that sets forth minimum standards for the protection of intellectual property rights including patents, copyrights, trademarks, and original industrial designs (WTO 2003a:1-8).

We turn first to the problem of counterfeit banknotes, and then to counterfeit goods and services. Our main interest, however, is in the latter.

Counterfeit Banknotes

The use of paper money dates as far back as the seventh century in China. However, it did not come into widespread use in Europe until the seventeenth century. The Bank of England, for instance, began issuing banknotes very shortly after its establishment in 1694 (Bank of England 2004:1).

At the same time, the Massachusetts Bay Colony began issuing notes to cover the cost of military expeditions. In 1739 a printing firm owned by Benjamin Franklin devised a scheme to deter counterfeiting by printing colonial notes with unique raised impressions of patterns cast from plant leaves. The dollar became the official money of the United States in 1785. Eighty years later the Secret Service was established to control counterfeit banknotes (BEP 2004:1-2).

The U.S. dollar serves as a reserve currency around the world. An estimated \$450 billion of the \$760 billion U.S. banknotes in circulation are held abroad. About one in 10,000 U.S. banknotes

circulating abroad are counterfeit which is approximately the same ratio of counterfeit to authentic (U.S. Treasury 2006:viii-ix). Counterfeiting is motivated by the clear economic gain but often is associated with other crimes including drug trafficking, illicit arms dealing, and terrorism (Secretary of the Treasury 2003:iii).

In addition to hand inspection of banknotes, various methods are used today to detect counterfeit notes including magnifying glasses and UV lights along with electronic devices with built-in detectors (Deutsche Bundesbank 2004:2). The problem is that counterfeiting banknotes has become much easier with the use of high-tech photographic and printing equipment (U.S. Secret Service 2003:1). To deter the use of computing technology for counterfeiting purposes, the Central Bank Counterfeit Deterrence Group, a working group of 27 central banks and note printing authorities, persuaded several leading hardware and software manufacturers to adopt a system that prevents personal computers and digital imaging tools from capturing or reproducing the images of protected banknotes (BIS 2004:1). Shortly after these restrictions became public, computer geeks began reporting several easy ways to circumvent them (EWF 2004:2), indicating that whatever counter-measures may be devised in the future likely will be only quick-fix remedies.

Important Distinctions: Counterfeit, Copycat, Overrun, Diverted

Hopkins et al (2003) define counterfeiting as “the knowing duplication of a product by a party who wishes to usurp the brand or trademark of another.” The same authors differentiate counterfeits from copycats, overruns, and diversions. A copycat is “a copy of a product in form or substance with no attempt to actually duplicate the brand name.” With a counterfeit product, deception

is intended; with a copycat, no deception is involved. An overrun is a product that has been ordered into production by the brand owners, has been overproduced, and sold directly by the manufacturer in the market, thereby violating the brand owner’s rights. With an overrun, the deception is that the seller is not the rightful owner of the trademark or brand name. Diverted products are ones that have been shipped into specific distribution channels and subsequently transferred into other distribution channels, thereby violating a contractual agreement (Hopkins et al 2003:9).

Counterfeit Products: Injustice and Risks

Counterfeiting violates the requirements set forth in the principle of commutative justice. Both parties to any marketplace exchange have two duties to one another under this principle: to exchange things of equal value and to impose equal burdens on one another. The equal-burdens requirement means that both parties agree knowingly and willingly that the money paid (buyer’s burden) and product given up (seller’s burden) are equal. The equal-value requirement means that both parties knowingly and willingly have reached agreement on the price that under normal circumstances means the price that other buyers and sellers would have agreed to under similar circumstances. Put more simply, the agreed price approximates the market price.

In a routine marketplace exchange, both parties must realize a gain in order for the exchange to take place. The gain for the buyer is that the product that is gotten is valued more than the money paid. For the seller, the money received is valued more than the product that is given up. The exchange in other words is a positive-sum transaction. Economists typically refer to these two gains as consumer surplus and producer surplus. We prefer “gain” to

“surplus” because the latter suggests something that is superfluous whereas the former points to something that is crucial. Simply put, there is no exchange unless both parties realize a gain.

There is nothing unjust about these gains, *per se*. Injustice arises when the exchange is a zero-sum transaction, as when neither party knows beforehand that the item sold is defective. The money-back guarantee is the usual remedy for this unintended injustice. The boycott is an example of a negative-sum action that is harmful to both the boycotters and a specific seller and is tolerated whenever there is a graver issue of justice at stake, as when the targeted producer operates a sweatshop. Third-party intervention, perhaps by a trusted religious leader, may be required in such cases to resolve the conflict between the parties involved before normal (positive-sum) market conditions are restored.

In an exchange involving a counterfeit product the buyer is unaware that the product is not genuine and through this deception overvalues the product, and is denied the gain that is rightfully his/hers. The seller in other words uses deception to enhance the gain that he/she is able to extract from the exchange, changing a transaction from a positive-sum exchange to a zero-sum exchange. This deception in turn imposes an injustice on the silent party to this exchange in that the dishonest seller's gain based on the use of a trademark or brand name without the rightful owner's permission deprives that owner of the gain that would have been gotten had the buyer bought the genuine product from the rightful owner instead of a counterfeit product from the dishonest seller. The loss involved in the passing of counterfeit banknotes is limited to the party who accepted the banknotes on the premise that they were genuine when in fact they were worthless.

Whether the counterfeiting involves banknotes or goods, deception is a

requirement for the transaction to reach completion. It follows that counterfeiting attacks the trust necessary for markets to function effectively, that is for both parties involved to feel confident that they will not be denied the gain which they expect and which set up the conditions making the exchange possible.

Quite apart from the issues surrounding the principle of commutative justice, other problems at times much more serious may follow from the use of a counterfeit product. For example, counterfeit bolts used in a construction project may shear off, collapsing a wall or roof because they cannot bear the load of the more expensive genuine bolts for which they have been substituted. Counterfeit pharmaceuticals that contain none of the active ingredients of the genuine products may delay recovery from an illness, disease, or injury for which they have been prescribed or in the extreme may even contribute directly to the user's death. In the case of an infectious disease, fake drugs may contribute to the spread of the disease.

By definition there are no net benefits from counterfeiting because the exchange invariably is a zero-sum transaction. What one party gains the other party necessarily loses. Even so, are there any circumstances under which the redistribution may be justified? The principle of the double effect is instructive here.

First, the good effect must outweigh the bad effect, lest more harm is done than good. The good effect relates to the gain gotten by the deceiver. The bad effect relates to the loss taken by the party who has been deceived plus any loss of trust that may block other market exchanges along with loss of sales of the genuine product endured by the authentic producer.

Second, the loss must not be deliberately intended. This condition could be met when the deceiver is in such dire circumstances and

so desperate that the only way the exchange can be executed is through deception, and the goods or services bought with the ill-gotten gains or gotten directly through deception are necessary to protect or sustain life. Pirating music, for example, could in the extreme be justifiable counterfeiting. It is difficult to imagine that counterfeiting pharmaceuticals ever is justified.

Third, and last, the principle of the double effect argues that the action taken in the exchange process must not be intrinsically immoral. Typically this would happen when coercion is applied by the deceiver, *forcing* the loss on the innocent party. To illustrate, a buyer is forced to purchase a counterfeit product or face the threat of a beating or worse. Clearly, the abused party in this case of extortion has been denied the freedom to withdraw from the transaction.

Since deception always is involved in counterfeit goods, and at times a loss of freedom as well, intervention inevitably is required. The principle of subsidiarity states that intervention should be left where possible in the hands of private groups such as trade associations and professional societies as long as they can address the matter satisfactorily and that public groups such as national governments should offer assistance to those private groups so that they are better able to carry out their responsibilities. Public groups should intervene directly only when private groups even with assistance are unable to address the matter satisfactorily. Given the cross-border activities of counterfeiters today, one is hard-pressed to argue that private groups can be effective especially when they operate alone. Minimally some type of partnership is needed between the private sector and the public sector on the one hand and national governments on the other.

Patents, copyrights, and trademarks afford the producer some protection against counterfeiting by affirming the right that the

producer has to the gain that is associated with whatever product or service he/she has created. Without that protection, which amounts to the right to sue in order to stop infringement and to retrieve damages for any gain that may have been lost to the counterfeiter, innovation would be stifled. Patents and copyright afford that protection for a fixed number of years. Trademarks, on the other hand, continue to protect without any such limit.

A very serious problem arises when proprietary rights clash with basic human need as in the case of patented pharmaceuticals that are life-sustaining but extremely costly. Indeed this clash often turns into a dilemma in that efforts to reduce the price of certain critical medicines to make them more affordable may erode the gain necessary to bring forth their development. For that reason, any balance that is struck between the incentive to research, develop, test, manufacture, and distribute a new drug and the affordability of that drug to those who need it unavoidably is uneasy, and must be re-struck from time to time by the human beings and the producing enterprises whose needs and incentives are at issue. The problem for the market system is that it is much better able to protect the gain of the producer than it is to address unmet human need. The answer lies in the hands of private groups and public bodies under the guidance of the principle of subsidiarity.

Five Critical Areas of Counterfeiting

This section addresses counterfeiting in five areas—parts and equipment, accessories, websites, cigarettes, and software—that were selected to be indicative of the length and breath of the counterfeiting practice. Counterfeit drugs are addressed separately in the next section.

Parts and Equipment. Counterfeit spare parts are a growing business, especially for

equipment that requires costly maintenance such as commercial and military aircraft (DeVale 1999:1). Bell Helicopter warns the public that an aircraft rebuilt or remanufactured around the recovered identification data plate is not an authentic Bell product. To help buyers identify what is counterfeit and what is authentic, Bell maintains and publishes a listing of destroyed aircraft and their serial numbers (Bell 2003:1). The U.S. Justice Department in 2002 indicted the operators of United Aircraft and Electronics for selling used flight-critical parts such as turbine blades for jet engines that were misrepresented as new (Gordon 2002:1).

Counterfeiting is a problem as well for automobile manufacturers. DaimlerChrysler maintains an email address for taking information on suspected counterfeit parts (DaimlerChrysler 2003:2). Counterfeit spare parts for cars made by Russian producer AvtoVAZ in 2003 were reported to constitute 30-40 percent of total in-country sales (Esmerk 2003:1). Belts, spark plugs, windshield-wiper blades, filters, gaskets, brake linings, voltage regulators, replacement engines for cars and trucks, and collectible cars are specific examples of counterfeiting in the automotive business (see “fakes” at Autocluster.com 2003). Estimated losses in the 1990s due to counterfeit auto parts range from more than \$1 billion per year in the United States to more than \$12 billion on a global basis (Autocluster.com 2003; IACC 2003a:3).

Accessories. Designer handbags, scarves, wallets, and sunglasses by such high-fashion manufacturers as Louis Vuitton, Gucci, Kate Spade, Christian Dior, and Burberry are favorite targets of counterfeiters. The price of a counterfeit good often is more than 80 percent below the price of the authentic product. This kind of counterfeiting, while clearly not as serious as counterfeiting drugs,

reduces the net worth of the authentic manufacturer by cheapening the brand name and allows the counterfeiter to profit from the marketing, advertising, and design development work of the authentic manufacturer (O’Brien 2003:1-2).

Websites. Counterfeit websites mimic the appearance of authentic websites, passing off information that is to varying degrees questionable or misleading. Fake websites are constructed on the foundation of the illusion of legitimacy. Examples include websites for Martin Luther King, the Makah American Indian Tribe, WTO, and a special website (*The Ed Report*) that mimics U.S. government reports (Piper 2002:1-6; WTO 2001:1). For WTO, fake websites make it more difficult for the public to gain access to authentic WTO documents and contribute to the image of WTO as lacking in transparency (WTO 1999:1).

Cigarettes. Nearly seven million metric tons of tobacco, valued at approximately \$20 billion, are grown annually around the world. More than five trillion cigarettes are manufactured every year, principally in Brazil, China, India, Turkey, and the United States. Hundreds of chemicals are used in the production of cigarettes to make the smoke easier to inhale and to reduce the amount of tobacco in each cigarette. Today manufacturers are using more reconstituted tobacco because it is easier to add chemicals and to include leaf stems and dust that in the past had been discarded. On a global basis, tobacco kills almost five million persons every year. If consumption continues to increase the World Health Organization estimated that the annual death toll would reach 8.4 million by 2020, with more than 70 percent of those deaths in developing countries (WHO 2003:102, WHO 2002:1).

In addition to counterfeit cigarettes there are two other types of contraband cigarettes: unaccounted exports and bootlegged

cigarettes. There are no official data on the volume of contraband cigarettes worldwide and therefore the only figures available are at best estimates. Unaccounted exports are exports that are not recorded as imports and are presumed to be smuggled. In 2001 unaccounted exports represent perhaps as much as three percent of world cigarette production (JTI 2003:1-2).

Bootlegging increasingly is directed by terrorist organizations and organized crime elements and is difficult to control (JTI 2003:1; US Senate 2003:1). It occurs when large quantities of cigarettes are purchased in a low-tax jurisdiction and shipped to a high-tax jurisdiction for resale, allowing the bootlegger to appropriate the difference. In California in 2002, the state tax on a pack of cigarettes was \$.87; in the neighboring state of Nevada, the tax was \$.35 (SBE 2002:7). On a million packs, the incentive for a bootlegger is \$520,000. Counterfeit cigarettes are manufactured in several different countries including Indonesia, Vietnam, Russia, Philippines, and United Arab Emirates. China is considered the primary source of counterfeit cigarettes (JTI 2003:1-2; BBC 2002:2). Counterfeiting also involves counterfeit cigarette tax stamps.

In London, counterfeit cigarettes are known to contain more tar, nicotine, and carbon monoxide than standard cigarettes and are being bought by 10-to-14-year olds because they are cheaper (BBC 2002:1-2). Thus the high taxes that have been imposed on retail cigarettes in order to discourage cigarette smoking have opened the doors to bootleggers and counterfeiters. And even though no reliable information on the extent of global counterfeiting is available, it is clear from the number of customs seizures in different countries and the volume of seized counterfeit cigarettes that this is a growing and lucrative international trade. The IACC has connected terrorist organizations with

other counterfeit items including jewelry, accessories, household products, and apparel (IACC 2003b:14).

Software. Global losses due to pirated software in 2007 were estimated at nearly \$48 billion or \$8 billion more than in 2006. A total of \$14.1 billion in losses in 2007 originated in the Asia-Pacific region. The heaviest losses in that region -- \$6.7 billion -- were reported for China. In Central and Eastern Europe losses totalled \$6.4 billion with Russia accounting for \$4.1 billion. The worst offending country was the United States where losses in 2007 were estimated at \$8.0 billion (BSA 2008:2,10-11).

The worldwide piracy rate in 2007--pirated software as a percent of total software installed -- was put at 38 percent. However, fully half of the 108 countries included in the Business Software Alliance study had piracy rates above 61 percent. The worst offending countries were Armenia, Bangladesh, Azerbaijan, Moldova, Zimbabwe, and Sri Lanka, all with rates of 90 percent or higher. Rates were 25 percent or lower in the United States, Luxembourg, New Zealand, Japan, Austria, Belgium, Denmark, Finland, Sweden, and Switzerland (BSA 2008:2,4).

BSA is a private group of more than 25 software manufacturers including Adobe, Apple, Borland, Cisco, Dell, HP, Microsoft, Symantec and others which has been established to fight the unauthorized copying or distribution of copyrighted software (BSA no date a). BSA operates a piracy hotline and offers rewards of up to \$1 million for qualifying piracy reports (BSA no date b).

Counterfeit Drugs

Worldwide pharmaceutical sales amounted to an estimated \$690 billion in 2007 (IMS 2007:1). The sheer volume of global production, along with the high prices of many prescription drugs and the ever-increasing dependence on medicines to

prevent, cure, and ameliorate human pain and suffering are powerful incentives to manufacture, distribute, and sell counterfeit drugs.

A counterfeit drug may contain no active ingredients, too much or too little active ingredients, the wrong or contaminated ingredients, or may be manufactured in the wrong dosage (FDA 2003:4). In 2003 FDA claimed that about 10 percent of the drugs in South East Asia are counterfeit. In China counterfeit drugs account for 50 percent of the product on the market. Deaths in China due to the use of fake drugs have been put at 192,000 though there is considerable doubt as to the accuracy of that figure. In underdeveloped countries the amount of counterfeit drugs is thought to be around 40 percent (FDA 2003:4; Forzley 2003:16). China and India are the main sources of the active ingredients used in the manufacture of counterfeit drugs worldwide. The trade in counterfeit drugs resembles narcotics trafficking in that the product is sourced in one country, formulated into tablets or capsules in another country, packaged in a third country, and transhipped through other countries on its way to its final destination (Glover 2001:89, 90).

Detection and Authentication

In this final section, due to the great harm to human health and well-being from the use of counterfeit drugs, attention is directed to the problem of detection and authentication: when is a product a counterfeit drug and when is it authentic? The problem is compounded because counterfeiters are able to circumvent anti-counterfeiting measures within 18-24 months. The equipment available today to counterfeiters makes it difficult even for the authentic manufacturer to detect fake drugs (FDA 2003:11,17).

Economic globalization and deregulation have created greater opportunities for

counterfeiters (Cohen 2003:4). According to the United States Customs Service, the overall volume of pharmaceuticals shipped by mail is “enormous” (Durant 2002:45). Roughly two million parcels containing FDA-regulated products enter the United States via international mail, most of which are released by the Customs Service to the addressee without review by FDA (Dingell 2001:8).

Even when a parcel is reviewed by FDA, detection is made more difficult due to the mingling of fake drugs with the authentic product thereby reducing the probability that random sampling will identify any fake drugs in the shipment (Christian 2001:93).

Regulatory agencies are just becoming aware of the problem of counterfeit drugs and the risk they pose to public health. The first study to compile information on the extent of this problem was published in 2003. Addressing this problem will require more regulatory oversight in which counterfeit goods are seen as a disease mechanism (see Forzley 2003:30). However, fewer than one-third of developing countries have fully functioning drug regulatory agencies. Ten to 20 percent of sampled drugs fail quality control tests in many developing countries. Poor manufacturing practices often result in toxic, sometimes lethal, products (WHO 2003a:2). FDA states that it costs between \$6,000 and \$15,000 to authenticate a box of ten drugs (Hubbard 2001:66).

The task at hand can be separated into four processes: prevent counterfeit drugs from entering the distribution network; improve the detection and authentication of drugs; reduce the risk of harm from using counterfeit drugs; avoid adding unnecessarily to the cost of producing authentic pharmaceuticals (FDA 2003:24). But the simple fact that in late 2003 the task force assigned to assist the lead agency responsible for drug safety in the United States (FDA) had prepared an *interim* report indicates that there is much to be done

to deal with the growing problem of counterfeit drugs.

Conclusion

No one knows for certain the extent of the counterfeit trade on a worldwide basis. One apparently reliable estimate puts the figure at five-to-seven percent of global trade.

Viewing the problem from the perspective of the counterfeiter, the reasons for uncertainty are obvious: the ill-gotten gains from counterfeiting are a substantial incentive and counterfeiters succeed only when they escape detection. Detection in turn is made even more difficult by public officials who turn a blind eye to the counterfeiting trade or themselves profit from the ill-gotten gains. Viewing the problem from the perspective of the buyer, the reasons are obvious: the lure of gains to be achieved by buying at lower prices at times driven by addiction as with cigarettes or other pressing need as with drugs. Further, terrorist organizations and crime syndicates are engaged in this trade and they are especially skilful in escaping detection and avoiding prosecution.

With greater economic globalization and further liberalization of trade relations between countries, there appears to be no end in sight to the growth of the counterfeit trade.

Selected References

Autocluster.com. (2003) *Parts Pirates*.
www.autocluster.com/sahistory/id230_m.htm

BIS (Bank for International Settlements). (2004) *Central Banks and Technology Industry Join to Combat Banknote Counterfeiting*.
www.bis.org/cgi-bin/print.cgi

Bank of England. (2004) *A Brief History of Banknotes*.
www.bankofengland.co.uk/banknotes/history.htm

BBC News. (2002) *Warning Over Fake Cigarettes*. July 11.

www.news.bbc.co.uk/1/hi/england/2123054.stm

Bell Helicopter. (2003) *Counterfeit Aircraft*.
www.bellhelicopter.textron.com/content/customerSupport/flightSatety/counterfeit.cfm

BEP (Bureau of Engraving and Printing). (2004) *The New Currency*.
www.moneyfactorycom/newmoney/main.cfm/currency/history

BSA. (2008) *Fifth Annual BSA and IDC Global Software Piracy Study*.
http://global.bsa.org/idcglobalstudy2007/studies/2007_global_piracy_study.pdf

BSA. (no date a) "About BSA & Members".
www.bsa.org/country/BSA%20and%20Members.aspx

BSA. (no date b) "Report Piracy".
www.bsa.org/country/Report%20Piracy.aspx

Christian, James. (2001) *Testimony before the Subcommittee on Oversight and Investigations*. Washington DC: House of Representatives, United States Congress, June 7.

Cohen, Mark Allen. (2003) *Statement: Roundtable Discussion before the Congressional Executive Commission on China*. Washington DC: United States Congress. February 3.

DaimlerChrysler. (2003) *Spare Parts*. London.

Deutsche Bundesbank. (2004) *Welcome to the Counterfeit Money Unit*.
www.bundesbank.de/bargeld/euro_falschg.en.php?pf=true

DeVale, John P. (1999) *Shoddy Spares: Customer Circumvention*. Spring.
www.ece.cmu.edu/~koopman/des_s99/spares_customer

Dingell, John. (2001) *Remarks before the Subcommittee on Oversight and Investigations*. Washington DC: House of

- Representatives, United States Congress. June 7.
- Durant, Elizabeth. (2002) *Statement*. Washington DC: Special Committee on Aging, United States Senate. July 9.
- Esmerk. (2003) *Russia: AvtoVAZ Changes Sales of Spare Parts*. September 23.
- EWf (Emulator World Forums). (2004) *Photoshop CS Was Built With Secret Anti-Counterfeiting Measure*. January 11. www.emulatorworld.com/forums/showthread.php?threadid=5165
- FDA. (2003) *FDA Counterfeit Drug Task Force Interim Report*. Rockville, MD: October.
- Forzley, Michele. (2003) *Counterfeit Goods and the Public's Health and Safety*. Washington DC: International Intellectual Property Institute. July.
- Glover, John D. (2001) *Prepared Statement and testimony before the Subcommittee on Oversight and Investigations*. Washington DC: House of Representatives, United States Congress. June 7.
- Gordon, John S. (2002) *Two Orange County Men Indicted in Fraud Scheme Involving Aircraft Parts*. California Central District: United States Attorney. April 4.
- Hopkins, David M.; Lewis Kontnik, and Mark Turnage. (2003) *Counterfeiting Exposed: Protecting Your Brand and Customers*, Hoboken, NJ: John Wiley and Sons.
- Hubbard, William K. (2001) *Prepared Statement*. Subcommittee on Oversight and Investigations, House of Representatives. Washington DC: United States Congress. June 7.
- IMS. (2007) "IMS Predicts 5 to 6 Percent Growth for Global Pharmaceutical Market in 2008, According to Analyst Forecast". www.imshealth.com/ims/portal/front/articleC/0,2777,6599_3655_82713022,00.html
- IACC. (2003a) *Facts on Fakes*. www.iacc.org Washington DC.
- IACC. (2003b) *International/Global Intellectual Property Theft: Links to Terrorism and Terrorist Organizations*. Washington DC.
- IACC. (2008) "Get Real – The Truth About Counterfeiting", www.iacc.org/counterfeiting/counterfeiting.php
- International Trademark Association. (1998) *The Economic Impact of Trademark Counterfeiting and Infringement*. New York: ITA. April.
- JTI (Japanese Tobacco International). (2003) *Facts and Figures About Contraband*. December.
- Lowe, Peter. (2003) *Email Message to Edward J. O'Boyle*. December 5.
- NAPM (National Association Purchasing Management). (2000) *Buyliner*. Southern Nevada. November.
- O'Brien, Diane. (2003) *When Imposters Knock Off Profits*. December 4. www.brandchannel/start.asp
- O'Connell, Thomas; Elizabeth Weiner; Hazel Bradford; Amy Borrus and Dorinda Elliott. (1985) *The Counterfeit Trade*. *Business Week*, December 16.
- OECD. (1998) *The Economic Impact of Counterfeiting*. Paris.
- OECD. (2008) *The Economic Impact of Counterfeiting and Piracy*, www.oecd.org/document/4/0,3343,en_2649_34173_40876868_1_1_1_1,00.html
- Pharmaceutical Security Institute. (2002) *About PSI*. www.psi-ncomabout.htm
- Piper, Paul S. (2002) "Web Hoaxes, Counterfeit Sites, and Other Spurious Information on the Internet", in Anne P. Mintz (Editor), *Web of Deception: Misinformation on the Internet*. Medford, NJ: CyberAge Books.
- Secretary of the Treasury. (2003) *The Use and Counterfeiting of United States Currency Abroad, Part 2*. Report to the

- Congress, March 2003. Washington DC: Congress.
- SBE (State of California Board of Equalization). *State Legislative Bill Analysis: SB 1849*. August.
- USDJ. (2002) *Press Release*. July 17. Washington DC
- U.S. Secret Service. (2003) *History of Counterfeiting*. Washington DC. www.secretservice.gov/counterfeit.shtml
- U.S. Senate. Kohl. (2003) *Hatch Introduces Bill to Halt Contraband Cigarette Trafficking Linked to Terrorist Funding*. June 3. Washington DC. www.senate.gov/~kohl/press/060303.html
- U.S. Treasury. (2006) *The Use and Counterfeiting of United States Currency Abroad, Part 3*, September. www.federalreserve.gov/boarddocs/rptcongress/counterfeit/counterfeit2006.pdf.
- WHO. (2002) *Illicit Tobacco Trade Contributes to Global Disease Burden*. Press Release/WHO 62, New York. July 4.
- WHO (2003) *Tobacco: The World Health Organization's Response*. www.who.int/whr/media_centre/factsheet2/en/print.html
- WTO. (2003a) *TRIPS: A More Detailed Overview of the TRIPS Agreement*. www.wto.org/english/tratop_e/trips_e/intel2_e.htm
- WTO. (2001) *Warning: Fake*. October 30. www.wto.org/english/news_e/news01_e/gattdotorg_e.htm
- WTO. (1999) *Press Release 151*. Nov 23. Geneva. www.wto.org/english/thewto_e/minist_e/min99_e/english/press_e/pres151_e.htm

Edward J. O'Boyle
 Mayo Research Institute
 West Monroe, Louisiana, USA
 edoboyle@earthlink.net

Criminal Justice: Comparative

Anne Cross

Introduction

Comparative criminal justice research presents and analyzes crime and justice data of different countries and cultures, both from an historical and contemporary perspective. The field examines social, political and global influences on the development and realities of criminal justice practice and identifies implications for public policy. Comparative approaches to legal and criminal justice systems provide important insights into criminal justice—and public policy in general. Since criminal justice systems are part of the social fabric and the bureaucracy, it can be difficult to analyze them objectively without a basis of comparison. In this regard, comparative criminal justice presents an opportunity for new insights into justice systems—both foreign and familiar.

To better understand a country's criminal justice system—and criminal justice in general—it is helpful to back away from analysis of a single system and benefit from having other cases with which to compare and contrast. As globalization brings the countries of the world into closer and more regular contact, comparative research is especially useful. Awareness of multiple models of criminal justice is particularly beneficial in our increasingly interconnected world wherein crime (and its causes) crosses national boundaries and often requires multinational cooperation on the part of law enforcement and criminal justice agencies (Rounds 2000).

The existence of multiple criminal justice systems worldwide also provides the basis of natural experiments in which variables can be isolated and studied. Cross-national comparisons can help researchers test theories about crime's causes, origins and distribution

(Reichel 2002). The development and growth of two international surveys – The United Nations Survey on Crime Trends and Operations of Criminal Justice Systems and the International Crime Victim Survey (ICVS) – have helped increase the collection and sharing of criminal justice data, as well as the precision of measurement. Along with adding to the knowledge base of the field, these surveys also embody some of the difficulties of comparative analysis, particularly in definition and measurement, which vary across cultures (Zvekic 1996).

The field of comparative criminal justice has grown and solidified since the 1970s when a critical mass of empirical studies first appeared. While there was not a tremendous overlap in the findings of the studies, a central tenet emerged from Western criminal justice researchers: that cross-cultural analyses are an effective means of understanding crime both at home and abroad.

A quick glance at crime statistics reveals that different countries have different realities in terms of crime and justice. For example, murder rates differ dramatically from country to country. Individuals are less likely to commit murder in Saudi Arabia than they are in Canada. Why might explain this? At first glance we might look to differences in the severity of punishments between Saudi Arabia and Canada. Saudi Arabia uses corporal punishments like public flogging and amputation that are unheard of in Canada. Some scholars have argued that the low rates of crime in Saudi Arabia are a result of Islamic law and its strict punishments (Souryal et al 1994; Souryal 1987). We might look to the influence of religion on the structure of each society. Criminal law in Saudi Arabia is developed around the religion of Islam and it forbids speaking out against Islam or the Saudi Royal family. In Canada, separation of religious faith and criminal

codes are taken for granted and there are accepted channels for expressing opposition to the government and dissent. In each society, there are basic understandings about what counts as a crime and how to best deal with criminal behavior. Defendants in Canada are offered more protections, compared with defendants in Saudi Arabia. Defendants in the Saudi system usually face strong cooperation between the judge and the investigator and a weak role for the defense attorney. Court procedures are more informal in Saudi courts and proceedings can be held in secret (Fairchild & Dammer 2001). We can look to the legal traditions and institutions of each country to help understand crime patterns. The different contexts of Canada and Saudi Arabia provide some clues to their very different crime rates.

At the end of our analysis (or even at the beginning) of criminal justice in Canada and Saudi Arabia it would be easy to conclude that the political and cultural backdrop of Saudi Arabia and the Canada are sufficiently different that it is not surprising that their crime and justice statistics would follow suit. As these examples show, Canada and Saudi Arabia are very different nations. Gaining an understanding the impact of different specific policies is difficult, since a firm basis for comparison is absent in these wildly divergent countries. But what happens when we compare the murder rates of relatively similar societies like Canada and the United States? While the difference in murder rates is not as extreme, neither is the cultural divide. How can we explain these differences? What insights can comparative criminal justice provide to better understand different societies and cultures?

Comparative Analysis as a Method

Comparative criminal justice offers a significant body of information about different countries. Beyond that, it offers a

method of investigation and analysis that has key advantages over other approaches (Archer & Gardner 1984; Ebbe 1996; Meyer 1972). Comparative research allows observers to examine or “test” theories and policies under different circumstances. Studies in comparative criminal justice tend to employ one of three strategies: single-nation studies (comparative in that they draw on other case studies), two-culture studies and more comprehensive studies directly examining three or more countries. Since criminal justice systems develop over time, many studies also apply historical analysis (Deflem & Swygart 2001).

While comparing countries with wildly divergent criminal justice systems can be interesting—for example, comparing Saudi Arabia with United States—researchers have often found it more fruitful to compare systems that more closely resemble one another in terms of structure and ideology. Comparing similar systems provides a more manageable and realistic basis of study and it also reaps results that are more easily applied to public policy concerns. While democracies like the U.S. and Canada are unlikely to import crime control methods such as public flogging, some research suggests that the British tendency towards shorter sentences and alternatives to prison could be usefully transplanted to the United States to deal with the problem of prison overcrowding.

Types of Criminal Justice Systems

There is consensus among comparative scholars that the criminal justice systems in the world can be meaningfully organized into four groups: (1) Common Law; (2) Civil Law; (3) Socialist; and (4) Islamic (Cole & Gertz 1981; Fairchild 1993).

Common Law Systems

The common law system originates in unwritten local customs and social norms.

King Henry II of England institutionalized common law in 12th Century England by bringing together local customs into a unified system of law shared by all of England. In Common Law systems, legal precedent is central. Custom, tradition and legal precedent is favored over written statutes (Ehrman 1976; Reichel 1999, 2002).

As a general rule countries that have been colonized by Britain use Common Law systems. The United States, Canada, Australia, New Zealand, South Africa, India, Sri Lanka, Malaysia, Brunei, Pakistan, Singapore, and Hong Kong can be considered common law countries (Fairchild & Dammer 2001; Rounds 2000).

Common Law systems feature an adversarial element in which lawyers are pitted against each other; interpreting and framing the facts of the case according to the interests of their clients. Common law systems rely on oral reports of evidence in the courtroom, putting the public trial—or at least the threat of it—at the center of the criminal justice system (Archer & Gartner 1984; Hirschel & Wakefield 1995).

Civil Law Systems

Contrasting common law societies' reliance on precedent, civil law societies put most of the emphasis on written laws and statutes. Written laws are at the center of decisions (Fennel et al 1995). Civil law is practiced throughout most of the European Union and also in Japan.

In terms of historical development, civil law developed out of the Roman law of emperor Justinian's Corpus Juris Civilis. The key difference between civil law and common law systems lies less in the actual codification of laws than in the overall approach to codes and statutes. In civil law countries, the primary source of law is legislation. The written law is not open to as much interpretation as it is under the common law

system. While fewer rights are extended to the accused in the civil law system, when compared to the common law system, protections of the accused are believed to be greater under civil law systems than they are under Islamic or socialist systems (David & Brierly 1978).

Islamic Law

Islamic law provides detailed sets of rules governing all aspects of individual and group behavior (Amin, 1985). In sharp contrast to common and civil law countries, Islam does not separate legal and religious authorities but unites both in Islam, which prescribes behavior in all spheres of life (David & Brierly 1978; Reichel 2002).

Islamic law is rooted in religious values and derives its premises from the Koran and related writings that govern all aspects of human conduct, providing detailed guidelines for everything from personal interactions to financial policy to criminal law. Crime control strategies are integrated with Muslim religious culture which provides instructions for dealing with deviance along with all areas of life (Adler 1983). Islamic systems derive all their procedures and practices from interpretation of the Koran. For example, instead of a separate, formalized criminal code in Saudi Arabia and other countries, there is a system of religious courts which rely on the Koran and other religious writings. Because religion plays a very central role in Islamic systems, most of these nations are considered theocracies, where a firm separation between legal rule and religious rule does not exist (Dien 2004; Souryal 1987).

Socialist Law Systems

Ideologically, socialist systems spring from the Marxist-Leninist tradition that views the criminal justice system as a means of training people to fulfill their responsibilities to the

state and to work for the common good. Socialist systems have existed in many places, most prominently the former Soviet Union, Africa, Cuba and Asia. Whereas civil and common law systems emphasize personal responsibility, socialist regimes emphasize citizens' responsibilities to the state. The underlying philosophy of socialism has an optimistic view of human nature, viewing it as perfectible and essentially good. Socialism holds that crime and delinquency is not a necessary feature of humans and society, but that these traits are a byproduct of capitalist exploitation. In other words, socialists believe that inequality and exploitative relationships make people delinquent. Fair distribution of material and social resources would eventually make crime unnecessary, socialists believe. Socialists believe that eventually all laws—including criminal laws—will not be necessary. Socialism is also characterized by extensive administrative law, where non-legal officials make most of the decisions (Sanders & Hamilton 1992).

Local Variations among Common Types

While the typology provides a useful framework, it is important to recognize the wide variety of local variation in criminal justice systems of the same basic type. Canadian justice, for example, places more emphasis upon the right to a fair trial, free from prejudicial publicity than does the United States. In Canada, the public and the media are usually banned from the courtroom and crime news is not aggressively reported. In England, there is more emphasis upon fairness in sentencing, and ensuring that the guilty don't go free when compared to the United States.

Each individual society also has unique traditions and legal histories that are resilient and influential. Many parts of Africa, for example, have reverted back to a tribal system because the imported system of

governance whose features were foreign to the native culture and historical tradition proved unworkable. Tribal justice systems tend to be characterized by arrest without trial and other arbitrary procedures (Thompson and Potter 1997). Other parts of the world, including Latin America, have seen their criminal justice systems struggle with the social changes and pressures brought by modernization (Shelly 1981). The growth of technology and industry, as well as profits of the drug trade, have brought rapid social change and have strained the capacity of the criminal justice system. The result in many Latin American countries has been a haphazard mishmash of legal practices (Mora 1996).

Comparative criminal justice research pays close attention to the variations that exist between nations because of their different backgrounds in terms of cultural values and political ideas. The typologies provide one manner of structuring comparative analysis, but other outlooks—particularly geo-historical approaches to comparative analysis have proven useful as well.

Hybrids and Misfits

Clashes between different criminal justice “types” in the same nation create scenarios not readily interpreted by way of the standard typological model. A prime example is found in the justice systems of many African nations. Some African countries have struggled to navigate contradictory influences of colonialism and traditionalism. Colonialism is, of course, the system under which the political and economic systems in Africa (and elsewhere) were controlled by Western countries. Traditionalism, on the other hand, promotes the values and ideas that have evolved over time in the region. During colonial rule, Western powers introduced values and criminal justice systems that were foreign in the African context and created

additional tensions and conflicts within the criminal justice system. For example, a study of criminal justice in Sierra Leone reveals several clashes between traditional and imported law (Thompson and Potter 1997). In Algeria, a French-imported system of law was uneasily combined with both Socialist doctrine and the religion of Islam (Adler 1983). In Tanzania, socialism was awkwardly mixed with traditional governance (Arthur 1996)

Elemental Analysis

For some analyses, the comparative approach is best focused on a specific aspect of the criminal justice system, such as courts, corrections or law enforcement. Law enforcement agencies have been especially well researched from a comparative perspective (Anderson 1989; Bayley 1985; Deflem, 1994; Mawby 1990). Law enforcement systems in the Western nations have been particularly well researched.

Several important comparative studies of law enforcement examine the function and organization of policing in different regions and countries (Huggins 1998; Bayley 1985). This approach to research has ferreted out important insights concerning the similarities and differences that exist between police systems across the world. In all countries, police systems are charged to maintain order and control crime. The manner by which law enforcement agencies carry out this charge varies and there is a great deal of diversity in the structure of law enforcement around the world. One commonality among law enforcement agencies is the challenge of juggling accountability with the unique powers afforded to most police forces, including that of arrest, seizure and in some cases, lethal force.

A common revelation in comparative studies of law enforcement is that the British patterns of law enforcement (inherited by the

American and other systems) are very different from those of other European nations. Outside of the U.K., most European nations organize law enforcement around military-like forces that operate from a centralized command. In contrast, the British and American systems are locally run and are more closely connected to the communities they police.

Outside of Western democracies, the most studied police organization is that of Japan. One of the reasons that policing in Japan has been interesting to researchers is that crime rates are lower there than in most other industrialized countries (See Bayley 1991; Ferdinand 1994; Miyazawa 1992; Steinhoff 1993). Some analysis suggests that the low crime rate may be connected to the comparatively high level of citizen involvement in the Japanese justice system (Ferdinand 1994; Westermann and Burfeind 1991). Citizen involvement is thought to encourage a high degree of overlap between the values of citizens and the direction of it legal system (Schneider 1992; Aldous & Leishman 1997).

Budding Democracies

Studies of emerging democracies highlight some of the challenges brought by political change. Often, increased crime occurs—at least temporarily—when repressive or totalitarian regimes are replaced by democratic ones. Since the 1990s, criminal justice researchers have examined police forces of countries transitioning from non-democratic regimes to democracies with democratic criminal justice organizations. The first major wave of research in this area examined nations that emerged from the former Soviet Union (Shelley 1996). Noteworthy studies have also been done on South Africa (Broden & Shearing 1993).

Law enforcement in Russia, for example, faced new and more vexing forms of criminal

behaviors when new freedoms were introduced. Money laundering and drug trafficking have risen markedly. South Africa also faced additional policing challenges when the Apartheid regime was lifted. Sudden migration occurred when the separatist rules governing where individuals could live were lifted. The resulting massive urbanization compounded problems of inequality and crime and also challenged the capabilities of police and other criminal justice agencies (Koelble 1999; Klug & Arup et al 2000).

The struggles involved in democratic transition are often experienced very acutely by police departments. Police in new democracies are charged with helping to breathe democracy into society at the street level while simultaneously facing a credibility gap there. They face the difficult task of creating a civil police force from one that was used for repressive and political purposes. In other cases, governments must create new forces from scratch. A successful transition to a democracy requires that violations be dealt with in a manner that protects the rights of suspects and other citizens. This task challenges even established police departments with strong track records and a legacy of citizen trust. For newly constituted police departments—which often face shortages in training, personnel and resources—democratic transition can be a very difficult task (Allen 1993).

A hangover effect is often left when non-democratic regimes are dismantled. It can impede the transition to democracy. Newly democratized criminal justice systems have frequently found it difficult to gain the trust of citizens. Long-held associations between police, prisons and repression can prove difficult to shake. These associations are often especially resilient for minorities and people who have experienced persecution and injustice.

In South Africa, for example, black populations tend to connect the criminal justice system with the repressive Apartheid regime. This, paired with the gap in job skills and educational attainment left by Apartheid, creates difficult challenges for democratic governance. Lingering perceptions about the police as a force of oppression can impede the operation of newly democratic police organizations. Creating a civilian police force that maintains order and controls crime in a democratic way has been an enduring challenge (Brogden & Shearing 1993).

Diversity Across Cultures

In addition to studying different institutional elements of the criminal justice system across political boundaries, comparative criminal justice can provide some general insights into the experiences of minority groups. Like crime, discrimination is universal but the form it takes varies by country. In all societies examined, members of minority groups are overrepresented in arrests and imprisonment and underrepresented among the ranks of workers and administrators in the criminal justice system (Tonry 1997; Killias 1997; McDonald 1997; Roberts & Doob 1997; Tsuda 1997).

In North America and Europe, racial and ethnic bias has been a well-documented feature of criminal justice. Minorities have a much higher rate of pre-trial detention and are more likely to receive prison sentences rather than probation for similar crimes. Research also suggests that minorities get inferior treatment (Killias 1997; McDonald 1997; Marshall 1997; Roberts and Doob 1997; Sampson & Lauritsen 1997; Tonry 1997). Minorities and immigrants are overrepresented in all stages of the criminal justice process. For example, Pierre Tournier (1997) conducted research in France and found that foreigners are overrepresented among suspected offenders and arrestees. A

large proportion of the suspected offenses are immigration-related.

Age and Gender

Age and gender play a determining role in criminal behavior in all cultures. Surprisingly, analyses of age and gender have not yet played a major role in comparative criminal justice studies (Curran and Cook 1993). This is unfortunate, since examination of the demographic characteristics of countries—particularly their age and sex structure—might go a long way towards explaining crime patterns, developing effective public policy and forecasting infrastructure needs.

While reliable comprehensive studies involving age and gender have not surfaced in large numbers in comparative criminal justice, a smattering of studies exist. Cross-cultural comparisons of youth cohorts have drawn on the connection of age and criminal behavior, noting implications for the needs of criminal justice systems (Cook & Davies 1999; Dobash & Dobash et al 1990; Pampel & Gartner 1995).

With respect to gender, scholars have undertaken interesting comparative analyses using data from the United Nations Crime Surveys. For example, in all countries surveyed between 1975 and 1985, men outnumber women by large margins at all stages of the criminal justice process, from suspicion to apprehension, prosecution, conviction, and imprisonment. Since the 1980s more women are present in the criminal justice system than ever before, although the number remains low compared to male representation (Harvey & Burnham et al 1992). While the reasons behind the changing gender trends are difficult to track empirically, scholars have identified changing social norms, additional opportunities for women to commit crime, and changes in women's expectations as primary causes.

Methodological Challenges

The difficulties involved in comparative criminal justice should be recognized, along with the limits of its application. Comparative criminal justice is based on the idea that we can advance knowledge by making systematic comparisons between nations. In the sense of being systematic, comparative criminal justice research differs from arm-chair comparative analysis, which can generate conclusions not supported by evidence. It also differs from traditional single-country studies that do not reap the advantages of comparison. Comparative criminal justice research has demonstrated that different political economies, different cultural backdrops and different contexts generate different outcomes. It has demonstrated that different historical circumstances generate different ideas about justice and different responses to crime. At the center of comparative research is this recognition that country-specific cultural, social, economic, and political pressures impact outcomes—a historically under-utilized concept in research and practice—and one that makes comparative researchers conservative about drawing hard and fast conclusions. What is successful in one country may not necessarily work in another because of the different social, economic and political contexts.

Despite the caution with which the field draws generalizations about crime and justice, several theories have emerged that provide insights for governance and public policy making.

General Theories

The following generalized theories have been put forward in the field of comparative criminal justice.

1. As a nation develops, awareness of crime increases and more crime is reported to the police.

2. With economic development, expectations of the police rise. People tend to expect that with economic development the police will become more effective at fighting crime.

3. Crime is linked with urbanization and overpopulation in urban areas (Archer & Gartner 1984).

4. As economic standards of living rise, victims become less careful about their belongings, and opportunities for committing crime expand.

5. Demographics shape a country's crime and justice profile. Countries with a large cohort of adolescents, for example, have more crime than countries with fewer adolescents.

6. Economic progress tends to bring rising expectations about the standard of living. These expectations can be unrealistic. People may feel the need to commit crimes to achieve them.

7. As societies become more complex, crime rates tend to rise. The spread and development of technology tends to make all nations increasingly similar and developing countries follow the same patterned increases in crime experienced by developing nations. This is the known as the modernization thesis (Shelley 1981).

8. As countries modernize, customs and norms that used to hold society together tend to break down causing informal mechanisms of social control to disintegrate and crime to rise.

Comparative Criminal Justice and Policy

While the field of comparative criminal justice is mature in the sense of having generated codifications, categories, methods and theories around which there is consensus, there is little evidence that the resulting scholarship has impacted public policy in a substantive manner. The character and quality of field's scholarship and research (and the summary of it presented above) bears the

biases of its intellectual creators and their followers who tend to be Caucasian Western men of middle-class or wealthier upbringing. The subjects of most interest have been political rather than social or economic and the resulting political studies have in most cases connected only superficially with related concerns such as the international political climate, distribution of natural and other economic resources, ethnic and cultural diversity, social history and ethics. Studies have been pursued from the top down, with judges, lawyers and legal documents taking precedence over analyses of lawbreakers, victims of crime, and family and community structure. These factors remain under-studied and not well understood. Courts (and police to a lesser extent) have taken center stage in the field's empirical development while corrections, victimology diversity, ethics, and rehabilitation have not been emphasized in the field's defining publications or textbooks. Economic development, foreign policy, and public education and literacy remain outside the paradigm.

Selected References

- Adler, F. (1983) *Nations Not Obsessed with Crime*. Littleton, Colo.: Rothman.
- Aldous, C. and F. Leishman. (1997) "Policing in Post-war Japan, Reform, Reversion and Reinvention", *International Journal of the Sociology of Law*, Volume 25, Number 2, pp. 135-54.
- Allen, G.F. (1993) Restructuring Justice in Russia, *Federal Probation*, Volume 57, pp. 54-8.
- Amin, S.A. (1985) *Islamic Law in the Contemporary World*. Glasgow, Royston.
- Anderson, M. (1989) *Policing the World*. Clarendon Press, Oxford.
- Archer, D. and R. Gartner. (1984) *Violence and Crime in Cross-National Perspective*. New Haven, Yale University Press.

- Arthur, J.A. (1996) "Crime and Penal Policy in the Socialist African Republic of Tanzania", *International Journal of Offender Therapy and Comparative Criminology*, Volume 40, pp. 157-73.
- Bayley, D.H. (1985) *Patterns of Policing*. New Brunswick, NJ: Rutgers University Press.
- Bayley, D.H. (1991) *Forces of Order, Policing Modern Japan*. Second Edition. Berkeley, CA: University of California Press.
- Brogden, M., and C.D. Shearing. (1993) *Policing for a New South Africa*. New York: Routledge.
- Cook, S., and S. Davies. (Editors) (1999) *Harsh Punishment, International Experiences of Women's Imprisonment*. Boston, Northeastern University Press.
- Curran, D.J., and S. Cook. (Editors.) (1993) Special Issue on: Crime and Justice in China and Japan. *Crime & Delinquency*, Volume 39, pp. 275-392.
- David, R., and J.C. Brierly. (1978) *Major Legal Systems in the World Today*. New York: Free Press.
- Deflem, M. (1994) "Law Enforcement in British Colonial Africa", *Police Studies*, Volume 17, Number 1, pp 45-68.
- Deflem, Mathieu and Amanda J. Swygart. (2001) "Comparative Criminal Justice", in T. DuPont-Morales, M. Hooper and J. Schmidt (Editors), *Handbook of Criminal Justice Administration*. New York, Marcel Dekker Publishers, 51-68.
- Dien, M. (2004) *Islamic Law, From Historical Foundations to Contemporary Practice*. Notre Dame: University of Notre Dame Press.
- Dobash, R.P.; R.E. Dobash; S. Ballintyne; K. Schumann; R. Kaulitzki and H.W. Guth. (1990) "Ignorance and Suspicion, Young People and Criminal Justice in Scotland and Germany", *British Journal of Criminology*, Volume 30, Number 3, pp. 306-20.
- Ebbe, O. (1996) (Editor) *Comparative and International Criminal Justice Systems*. Boston: Butterworth Heinemann.
- Ehrman, H. (1976) *Comparative Legal Cultures*. Englewood Cliffs, NJ: Prentice Hall.
- Fairchild, E. (1993) *Comparative Criminal Justice Systems*. Belmont, CA: Wadsworth.
- Fairchild, E. and H. Dammer. (2001) *Comparative Criminal Justice Systems*. Belmont, CA: Wadsworth.
- Fennell, P.; C. Harding; N. Jorg and B. Swart. (1995) *Criminal Justice in Europe, A Comparative Study*. New York: Oxford University Press.
- Ferdinand, T. N. (1994) "Japanese Criminal Justice", *International Criminal Justice Review*, 4, 72-9.
- Harvey L.; R.W. Burnham; K. Kendall and K. Pease. (1992) "Gender Differences in Criminal Justice, An International Comparison", *British Journal of Criminology*, Volume 32, Number 2, pp. 208-17.
- Hirschel, J. D., and W. Wakefield. (1995) *Criminal Justice in England and the United States*. London: Praeger.
- Huggins, M.K. (1998) *Political Policing, The United States and Latin America*, Durham, NC: Duke University Press.
- Killias, M. (1997) "Immigrants, Crime and Criminal Justice in Switzerland", *Crime and Justice*, Volume 21, pp. 375-405.
- Klug, H.; C. Arup; M. Chanock and P. O'Malley. (2000) *Constitution Democracy, Law, Globalism and South Africa's Political Reconstruction*. Cambridge, U.K.: Cambridge University Press.
- Koelble, T. (1999) *The Global Economy and Democracy in South Africa*. New Brunswick: Rutgers University Press.

- Lopez-Ray, M. (1970) *Crime, An Analytical Appraisal*. New York: Praeger.
- Marshall, I. H. (1997) *Minorities, Migrants, and Crime, Diversity and Similarity across Europe and the United States*. Thousand Oaks, CA: Sage.
- Mawby, R.I. (1990) *Comparative Policing Issues, The British and American Experience*. London, Unwin Hyman Ltd.
- McDonald, W.F. (1997) "Crime and illegal immigration", *National Institute of Justice Journal*, Volume 232, pp. 2-10.
- Meyer, J.C. Jr. (1972) "Methodological Issues in Comparative Criminal Justice Research", *Criminology*, Volume 10, Number 3, pp. 295-313.
- Miyazawa, S. (1992) *Policing in Japan*. Albany, NY, State University of New York Press.
- Mora, F.O. (1996) "Victims of the Balloon Effect, Drug Trafficking and U. S. Policy in Brazil and the Southern Cone of Latin America", *Journal of Social, Political and Economic Studies*, Volume 21, Number 2, pp. 115-40.
- Pampel, F.C. and R. Gartner. (1995) "Age Structure, Socio-Political Institutions, and National Homicide Rates", *European Sociological Review*, Volume 11, Number 3, pp. 243-60.
- Reichel, P. (1999) *Comparative Criminal Justice Systems*. Englewood Cliffs NJ, Prentice Hall.
- Reichel, P.L. (2002) *Comparative Criminal Justice Systems*. Upper Saddle River NJ, Pearson Education.
- Roberts, J.V., and A.N. Doob. (1997) "Race, Ethnicity, and Criminal Justice in Canada", *Crime and Justice*, Volume 21, pp. 429-522.
- Rounds, D. (2000) *Comparative Criminal Justice*. Boston: Allyn & Bacon.
- Shelley, L. (1981) *Crime and Modernization*. Carbondale: SIU Press.
- Sampson, R.J. and J.L. Lauritsen. (1997) "Racial and Ethnic Disparities in Crime and Criminal Justice in the United States.", in M. Tonry (Editor), *Ethnicity, Crime and Immigration, Comparative and Cross-National Perspectives*. Chicago, Chicago University Press.
- Sanders, J., and V.L. Hamilton. (1992) "Legal Cultures and Punishment Repertoires in Japan, Russia and the United States", *Law and Society Review*, Volume 26, Number 1, pp. 117-38.
- Schneider, H.J. (1992) "Crime and its Control in Japan and in the Federal Republic of Germany", *International Journal of Offender Therapy*, Volume 36, Number 4, pp. 307-21.
- Shelley, L. (1996) *Policing Soviet Society*, New York, Routledge.
- Souryal, S.S. (1987) "The Religionization of a Society, The Continuing Application of Shariah Law in Saudi Arabia." *Journal for the Scientific Study of Religion*, Volume 26, pp. 249-65.
- Souryal, S.S.; D.W. Potts and A.I. Alobied. (1994) "The Penalty of Hand Amputation for Theft in Islamic Justice", *Journal of Criminal Justice*, 22, 240-65.
- Steinhoff, P.G. (1993) "Pursuing the Japanese police", *Law and Society Review*, Volume 27, pp. 827-50.
- Thompson, B. and G. Potter. (1997) "Governmental Corruption in Africa, Sierra Leone as a Case Study", *Crime, Law and Social Change*, Volume 28, pp. 137-54.
- Tonry, M. (1997) "Introduction", *Ethnicity, Crime, and Immigration, Comparative and Cross-National Perspectives*. Ed. by M. Tonry, Chicago, Chicago University Press.
- Tournier, P. (1997) "Nationality, Crime, and Criminal Justice in France", in M. Tonry (Editor), *Ethnicity, Crime, and Immigration, Comparative and Cross-*

National Perspectives. Chicago: Chicago University Press.

Tsuda, M. (1997) "Human Rights Problems of Foreigners in Japan's Criminal Justice System", *Migration World Magazine*, Volume 25, pp. 22-5.

Westerman, T.D. and J.W. Burfeind. (1991) *Crime and Justice in Two Societies, Japan and the United States*. Pacific Grove CA: Brooks/Cole.

Zvekic, U. (1996) "The International Crime Victim Survey, Issues of Comparative Advantages and Disadvantages", *International Criminal Justice Review*, Volume 6, pp1-21.

Websites

Cecil Greek's Criminal Justice Resources.

www.criminology.fsu.edu/cjlinks/world.html

INTERPOL. (International Criminal Police Organization) www.interpol.int/default.asp

UNCJIN. (United Nations Crime and Justice Network) www.uncjin.org

United Nations Survey on Crime Trends. www.uncjin.org/Statistics/statistics.html

World Factbook of Criminal Justice. www.ojp.usdoj.gov/bjs/abstract/wfcj.htm.

Anne Cross

School of Law Enforcement & Criminal Justice, Metropolitan State University

Minnesota, USA

anne.cross@metrostate.edu

Criminal Justice: Punishment & Retribution

Mark D. White

Introduction

Criminal punishment represents a genuine quandary: it is omnipresent, existing in almost all societies through recorded history, but at the same time it has proven to be notoriously difficult to justify using the concepts of legal and moral philosophy. Punishment, by popular definition, “must involve pain or other consequences normally considered unpleasant” (see Hart 1968, Flew 1954 & Benn 1958, for seminal definitions of punishment). The intentional infliction of such treatment on persons is *prima facie* immoral, and therefore demands strong arguments to justify it.

This article will focus on the two primary justifications of punishment—deterrence and retributivism—as well as hybrid or mixed theories which combine the two. Briefly, deterrence justifies punishment in terms of its consequences, specifically the increase in well-being resulting from a decrease in criminal activity. Retributivism, on the other hand, justifies punishment in terms of the desert of the criminal, who is punished because of wrongdoing, regardless of any consequences. Advocates of mixed theories of punishment maintain that the general practice of punishment is justified by deterrence, but the actual punishment of individuals is constrained by retributive concerns. Other justifications of punishment include rehabilitation, incapacitation, moral education and community outrage; often these are tied to either deterrence or retributivism as elaborations on the two basic themes. (For essential work on punishment, see Action 1969; Duff & Garland 1994; Simmons *et al* 1995).

Deterrence

Origins and Central Ideas

The explicit reference to deterrence as general justification of punishment, and determination of specific penalties, is usually traced to Jeremy Bentham (1781) (drawing upon the early criminology of Beccaria 1764), and is a straightforward extension of utilitarian ethics to the problem of punishment. (See Tunick 1992:69-76, for a summary of Bentham’s theory of punishment.) Under deterrence, punishment is justified if it results in greater societal benefits than harms (or, more precisely, if it generates greater net benefits than any alternative). It treats punishment as a “bad” that creates disutility, and is only permitted if that disutility is overwhelmed by the benefits due to lower harm from crime: “All punishment in itself is evil. Upon the principle of utility, if it ought at all to be admitted, it ought only to be admitted in as far as it promises to exclude some greater evil” (Bentham 1781:170). Punishment thus has no intrinsic value; its value is purely instrumental and contingent on its effects.

Deterrence is inherently forward-looking, justifying punishment not in terms of something owed to or deserved by a particular offender for harms done, but instead in terms of improving future well-being by lessening crime. This implies that punishment is recommended according to its efficacy, and lack thereof can limit its appropriate application. Bentham lists a number of factors that render “cases unmeet for punishment” (1781:170), such as when punishment is “*groundless*: where there is no mischief for it to prevent...*inefficacious*: where it cannot act so as to prevent the mischief...*unprofitable*, or too *expensive*: where the mischief it would produce would be greater than what it prevented...[or] *needless*: where the mischief may be prevented, or cease of itself, without it: that

is, at a cheaper rate” (1781:171, original emphasis).

The fact that Bentham cited expense in the passage above demonstrates that the consequentialist outlook of deterrence places unique emphasis on the costs of punishment, such as maintaining physical facilities for imprisonment, staffing prisons, providing food and medical care for prisoners, etc. Given the emphasis on costs, deterrence also informs broader decisions in criminal justice, such as the allocation of scarce resources among detection, apprehension, prosecution, and punishment, and explains the existence of trade-offs made in the criminal justice system, such as plea bargains (in which one suspect is given a reduced sentence in exchange for information that can help prosecute another criminal), prosecutorial discretion, and early release or parole.

The Economic Approach to Crime

Deterrence theory, to be made operational, also requires a theory of human behavior on which to base the design of optimal deterrent punishments. In light of this, it is appropriate that economists have become the torchbearers of a deterrence approach to punishment, starting with Becker’s (1968) seminal work on the economics of crime. Becker applied the economic model of rational choice and the efficiency-maximization paradigm of welfare economics to the study of criminal behavior and punishment, resulting in a formal, mathematical representation of deterrence theory that verified many of Bentham’s intuitions, and also derived original, novel extensions.

A significant result from Becker (1968) is that, assuming equivalent amounts of deterrence, fines should be used as often as possible to punish crimes, because the resource costs of collecting fines is trivial and they raise revenue for the government. In economic terms, punishing by fines is merely

a transfer of wealth from offender to the state, with only minor deadweight loss from the collection process. This is in stark contrast to imprisonment, which involves nontrivial (often enormous) resource costs on the part of the state, with no offsetting revenue stream at all, resulting in a significant net loss to society. Imprisonment may still be welfare-improving if its deterrent effect is substantial enough to outweigh the costs, but if fines can provide the same deterrence, then social costs will be lowered by using fines. (Note that this assumes that a fine can be designed to provide the same degree of deterrence as a given prison sentence, and that offenders have the resources to pay the fines.)

Another insight of Becker’s, which has been elaborated on extensively in the literature (e.g. Polinsky & Shavell 1979), is the trade-off between the probability and magnitude of punishments. (Strictly speaking, the probability of punishing an offender is not strictly a function of punishment, but also of detection, apprehension, and prosecution, and therefore this refers to a broader criminal justice outlook than punishment alone.) The general idea is that the probability and magnitude of punishment must be chosen optimally to minimize the costs of crime, by providing efficient deterrence. If we assume combinations of probability and magnitude that provide equivalent deterrence, then the only consideration is cost. As Becker showed, fines involve much lower costs than imprisonment, and also much lower than the resource cost of apprehension, etc., which affect probability. So the lowest-cost option is to increase the fine as much as possible and reduce probability; the extra deterrence provided by the higher fine will enable the authorities to reduce the resources devoted to apprehension, lowering total costs. If imprisonment is chosen, then the problem is more complicated since both probability and magnitude of punishment are costly; also, the

deterrent effect of probability depends on offender's individual levels of risk aversion, which will affect the optimal probability/magnitude combination (Polinsky and Shavell 1999).

One aspect of classical deterrence theory (Bentham 1781:14) that has been particularly highlighted by economic analysis is proportional sanctions, which produce marginal deterrence (Friedman and Sjoström 1993). These terms refer to penalties being set in proportion to the seriousness of the crimes, such as murder being punished more harshly than assault, or grand larceny carrying a higher sentence than petty theft. The economic justification for this is that the lower penalty for lesser crimes will provide criminals with incentive to "substitute" the less harmful crime for a more harmful one. For instance, a higher penalty for felony murder than for theft will give thieves incentive not to kill witnesses—if the penalties were the same, witnesses would be killed more often because they enhance the thief's chances of escaping, without increasing the penalty if caught.

However, disproportionate punishment is also recommended by economic analysis to offset uncertain apprehension of criminals. If the probability of apprehension is only 25%, for instance, then the effectiveness of a given penalty will be diminished by approximately 75% because the criminal will discount the penalty by the probability of evading apprehension. Assuming risk neutrality, a crime that would be deterred by a fine of \$1000 (with certainty) would then have to be punished by a \$4000 fine (with 25% probability). However, if more serious crimes are solved more often (higher probability), then their penalties can be discounted less, and marginal deterrence could be compromised.

Criticisms

Since deterrence is derived from utilitarianism, the standard criticisms thereof apply to deterrence as well. (See Smart and Williams 1973 & Scheffler 1988 for general critiques of utilitarianism and consequentialism.) For instance, Nozick points out that, due to its utilitarian origins, deterrence "equates the happiness the criminal's punishment causes him with the unhappiness a crime causes its victim" (1974:61) by failing to make a moral distinction between utility rightfully and wrongfully gained. This aspect of deterrence is highlighted in the economic analysis of criminal punishment, which counts the criminal's disutility in the social welfare function that optimal punishments are designed to maximize.

More specifically, the forward-looking nature of deterrence ignores the wrong performed by the offender and the punishment that the criminal deserves for it. Under deterrence, punishment of a specific criminal is not focused on him, but intended instead to provide incentive, in the form of a credible signal, for other potential offenders to reduce their criminal activity. For this reason, critics contend that deterrence may well be served by falsely convicting and punishing an innocent person; if the subterfuge is successfully concealed, the message sent to potential offenders may be the same. Defenders of deterrence claim either that such a scenario is unrealistic (and therefore irrelevant), or that such an institution of false persecution would itself lower total utility sufficiently to render it infeasible under utilitarian standards. But these responses are themselves contingent, and do not answer the charge that *if* such a ploy were possible, then deterrence would demand its practice.

To advocates of deterrence, punishment is only permitted if it is effective, since it is

prima facie bad without additional consequentialist justification. If punishment is too costly, if it has questionable deterrent effects, or if it can be compromised in order to secure the conviction of another suspect, then in these cases the deterrence approach requires that punishment be reconsidered, and possibly lessened or cancelled altogether. To critics, this is another instance of deterrence not recognizing the inherent wrong performed by the criminal, which requires punishment as a matter of justice and right, rather than contingent matters of maximizing well-being or utility.

Retributivism

Origins and Central Ideas

Though elements of the retributive viewpoint can be found throughout the history of philosophy, Immanuel Kant (1797) is regarded as the first significant retributive thinker; Hegel (1821) and Ross (1930) are also influential early retributivists. Retributivism can be considered a deontological approach to punishment, justifying its practice not in terms of consequences but instead with reference to the wrong performed by the criminal. To retributivists, punishment is given out as matter of justice rather than expediency, or promoting the right versus maximizing the good. (Nonetheless, Moore 1993 has suggested that retributivism can be conceptualized as a consequentialist theory in which punishment is an intrinsic good to be maximized; see Dolinko 1997 for a criticism of this theory, and below for other teleological variants of retributivism.)

There are various types of retributivism, with most being framed as a response to the criticisms of deterrence. One such criticism, described earlier, is that deterrence de-emphasizes the wrongful act and the criminal himself, and punishes him based solely on the

effects of doing so on overall welfare. Most retributivists do not accept expense, lack of deterrent power, or bargaining with one criminal to help convict another, as conclusive reasons to reduce punishment, and some would deny any relevance to such considerations: “woe to him who crawls through the windings of eudaemonism in order to discover something that releases the criminal from punishment or even reduces its amount by the advantage it promises” (Kant 1797:331).

Understood more generally, retributivists charge that the goal of deterrence leads to using convicted persons, both innocent and guilty, as means to an end, namely deterrence and general utility. Such treatment violates the deontological ideal (itself derived from Kantian ethics) of respecting the dignity of persons; Kant writes that state punishment “can never be inflicted merely as means to promote some other good for the criminal himself or for civil society. It must always be inflicted upon him only *because he has committed a crime*. For a human being can never be treated merely as a means to the purposes of another” (1797:331). This ideal is often invoked when addressing charges that deterrence sanctions the deliberate punishment of innocent persons; again, Kant writes that the criminal “must previously have been found *punishable* before any thought can be given to drawing from his punishment something of use for himself or his fellow citizens” (1797:331). To retributivists, even punishment of the truly guilty treats them as means to an end if the only reason they are punished is to deter future crimes, rather than to punish them personally. (Many other retributivists emphasize the respect given to persons under retributivism, including Hegel 1821; Morris 1968; Murphy 1973.)

Types of Retributivism

There are two broad varieties of retributivism, each of which can be justified in various ways. *Negative retributivism* places constraints on permitted punishment: the innocent must never be punished, and the guilty must never be punished excessively, or in a manner disproportionately severe relative to their crimes. *Positive retributivism* goes further by placing affirmative requirements on punishment: not only must the innocent not be punished, but the guilty must be punished. Furthermore, not only must the guilty not be punished excessively, they also must not be punished too lightly. (The distinction between positive and negative retribution is usually credited to Mackie 1985:207-208.)

Negative retributivism is less controversial, since it can be seen simply as a constraint on deterrent policies, particularly punishment of the innocent. But it also implies that the guilty cannot be punished disproportionately severely, which does interfere with flexibility of penalties often required by deterrence, especially as analyzed by economists. Recall that the economic approach to crime recommends the penalties and the probability of their imposition should be adjusted to achieve the lowest cost of a given level of deterrence. If we assume that, with certainty of punishment, the penalty would be set roughly to the “just” level (which is by no means certain), that any diminution in the likelihood of punishment would require an increase in the severity of punishment, which is forbidden by negative retributivism. Therefore, even this milder form of retributivism is at odds with the economic approach to punishment. (For discussion of retributivism within economic models of crime and punishment, see Wittman 1974; Posner 1980; Avio 1990,1993; Kaplow & Shavell 2002.) Plus, some have expressed doubts whether negative (or “weak”) retributivism should even be

called retributivism, since it does not mandate punishment as a matter of justice or right, but merely forbids punishment when it violates justice or right (Cottingham 1979:240-241). As described above, in this sense it is merely a constraint on nonretributivist policies; such hybrid theories of punishment are discussed below.

Positive retributivism, however, does require punishment of the guilty as a matter of justice and right, as well as forbidding punishment of the innocent, and also requires that the guilty are punished no less (or more) than their due. Positive retributivism normally has some connection to the *lex talionis*, the Biblical prescription of “an eye for an eye” and revisiting a wrongdoer’s crime upon him. Kant asks “what kind and what amount of punishment is it that public justice makes its principle and measure? None other than the principle of equality... Accordingly, whatever undeserved evil you inflict upon another within the people, that you inflict upon yourself.... But only the *law of retribution* (*ius talionis*)... can specify definitely the quality and the quantity of punishment” (1797:332). Few modern retributivists refer to the *lex talionis* directly, and some dispute its relevance to retributivism altogether (Davis 1986). Nonetheless, compared to negative retributivism, positive retributivism contrasts more directly with deterrence-based justifications of punishment because it specifies more precisely who must and must not be punished and how much they should be punished, regardless of the effect on deterrence or social welfare. These demands are particularly troubling given the fact of economic scarcity: no society can devote the resources necessary to apprehending and punishing *all* offenders without neglecting other essential human needs (Avio 1990).

The main problem with positive retributivism rests with its justification—it is

easy to say why the innocent must not be punished, but it is much more difficult to argue why the guilty *must* be punished (without resorting to deterrence as a rationale). Murphy (1971) distinguishes between those who believe that the guilty deserve punishment as a primitive axiomatic ideal, and those who base their justification of punishment on a “general theory of political obligation” (166). The former, named “intrinsic retributivists” by Honderich (1984:212), hold punishment of the guilty to be intrinsically good, and not contingent on any other normative qualities. (See Davis 1972 for a defense of the intrinsic desert position against criticisms from Honderich and others.)

Most retributivists adopt the second approach identified by Murphy, and justify punishment by reference to a broader political obligation on the part of citizens or the state. Though there are subtle differences, most of these justifications represent a sense of reciprocity, that the guilty person “owes” society a “debt” (also justifying the term “retributivism” or “retribution”; see Cottingham 1979). Some, such as von Hirsch (1976), Morris (1968), Murphy (1973), Finnis (1972), Sher (1987), Rawls (1964), and Kant himself (according to Murphy 1972, 1973), claim that punishment helps balance the benefits and burdens of obeying the law: criminal gain unfair advantage over other citizens who accept the restrictions of obeying the law. Others, such as Hegel, Cooper (1971), and Hampton (in Murphy & Hampton 1989), believe that punishment “annuls” the original crime (in a parallel to compensation in tort law): Hegel holds that the purpose of punishment “is to annul the crime, which otherwise would have been held valid, and to restore the right” (1821:69). (For discussion of the various reciprocity-based justifications of positive retributivism, see Ten 1987:ch.3, 1990; Braithwaite and Pettit

1990:ch.8; Dolinko 1991; Tunick 1992:ch.3; Duff 2001:ch.1.)

Often, these scholars emphasize that the criminal “has brought the punishment upon himself”: he “has chosen to be punished” (Morris 1968:36), because he freely chose an action (ideally) in full knowledge of the consequences. This is often seen by critics as nonsensical, but as Kant explains, “no one suffers punishment because he has willed *it* but because he has willed a *punishable action*; for it is no punishment if what is done to someone is what he wills, and it is impossible to *will* to be punished” (1797, 335). Even a murderer wants other murderers punished (if only to protect his own life), but to be consistent (and thereby rational), he must acknowledge that his own punishment is justly applied as well. Some scholars term this a “right” to punishment, language which may be (even) more problematic; see Hegel (1821:70-71) and Bosanquet (1965:201-211) for early proponents for this view, and Deigh (1984) for critique.

Other common justifications for positive retributivism, not based on reciprocity, include denunciation or expression of condemnation (Denning 1984, Feinberg 1965, von Hirsch 1985) and moral education and communication (Hampton 1984; Duff 1986, 2001). These are distinct from the reciprocity-based justifications in that they do not emphasize the suffering of the guilty, but instead some benefit arising from the punishment. Therefore, Nozick refers to such variants as *teleological retributivism* (1981:371): these theories are still distinguished from deterrence due to their strong link back to the particular criminal or crime, rather than broad, future benefits to society as a whole.

Aside from the issue of justification, positive retributivism also faces the difficult issues of whom to punish and how to punish proportionately to the crime (Braithwaite and

Pettit 1990:166-180). The question of whom to punish deals with broad concerns: moral issues such as responsibility and intentionality, as well as legal issues such as justification, excuse, and mercy. The question of how to punish is more technical: how exactly does the state ensure that the punishment is proportional to, or “fits”, the crime? The *lex talionis* demands the punishment “equal” the crime, but this is obviously impossible in many cases, such as rape, which Kant realized: “but what is to be done in the case of crimes that cannot be punished by a return for them because this would be either impossible or itself a punishable crime against *humanity* as such?” (1797:363). In answer, he moderates the *lex talionis*: “the only time a criminal cannot complain that a wrong is done him is when he brings his misdeed back upon himself, and what is done to him in accordance with penal law is what he has perpetrated on others, *if not in terms of its letter at least in terms of its spirit*” (1797:63, emphasis added).

“In terms of its spirit”, most retributivists recommend only proportionality, but this does not solve the problem so much as it redefines it. Card (1975) and von Hirsch (1976) are leaders in this area, and both are criticized in Bedau (1978) (with a rejoinder in von Hirsch 1978); Kleinig (1973), Nozick (1981, 363-365), and Davis (1983) also contribute to this question, the last building on a ranking process derived from Mabbott (1939) (and criticized by Benn 1958); and some (such as Wertheimer 1975) say the task of assigning a consistent scale of punishments based on retributivism (or deterrence, for that matter) is impossible. It would seem that deterrence advocates have a clear advantage in the determination of penalties, since their goal is to find the “best” (optimally deterrent) punishments rather than the elusive “right” (just) penalties sought by retributivists, but the difficulties with utilitarian calculation and

predictive models of rational choice cast any advantage into question.

Hybrid Theories of Punishment

In the last several decades, scholars have begun to suggest hybrid theories of punishment that attempt to combine deterrence and (negative) retributivism in a mutually consistent fashion. The most prominent proponent of such a system is H.L.A. Hart (1968), who distinguished between the “general justifying aim” of punishment and the “distribution” of penalties, or between why we punish and how we punish. (See also Benn 1958 for this distinction.) In this way, “it is perfectly consistent to assert *both* that the General Justifying Aim of the practice of punishment is its beneficial consequences *and* that the pursuit of this General Aim should be qualified or restricted out of deference to principles of Distribution which require that punishment should be only of an offender for an offence” (Hart 1968:9). (See also Ross 1930:61-64, Mabbott 1939, for early reference to this idea, and Rawls 1955, for a defense of this idea from the viewpoint of rule utilitarianism.) In the end, this is essentially the deterrence rationale constrained by negative retributivism, as mentioned above.

While some modern retributivists (such as von Hirsch 1985) have proposed or defended such a system, the most surprising source for this hybrid theory of punishment is Immanuel Kant himself, often considered the paradigmatic (positive) retributivist. But recently, scholars such as Byrd (1989), Murphy (1987), Scheid (1983), and Hill (1999) have defended a revised interpretation of Kant’s statements on punishment; they highlight many instances in his writings—outside *The Metaphysics of Morals* (1797), which contains what is considered his canonical work on punishment—in which he

seems to have defended deterrence as the overall goal of the system of punishment, while requiring that retributivist principles govern individual penalties (corresponding to Hart's general justifying aim and distribution, respectively).

While some, such as Wertheimer (1976), hold that a system of deterrent punishment requires retributivist distribution of penalties to render threats of punishment credible (similar to the rule-utilitarian argument in Rawls 1955), others argue that such a hybrid theory of punishment promotes neither deterrence nor retributivism. Braithwaite and Pettit (1990:166-168) maintain that a truly deterrent system of punishment must adhere to deterrent principles at every level, general justifying aim *and* distribution: whenever retributivist distribution of penalties would differ from those recommended by deterrence, deterrence is weakened. Goldman (1979) writes that a hybrid system of punishment may successfully prohibit punishment of the innocent, but cannot prohibit excessive punishment of the guilty, which is necessitated by the deterrence goal (see above). (This point was also made by Nozick 1974:60-61, but not explicitly in terms of a hybrid system of punishment.) Avio (1993) generalizes this inconsistency, pointing out deterrence and retributivism often demand different punishments, and, more fundamentally, "the interdependence of punishment as *threat* and punishment as *executed* requires a theory which addresses this interdependence" (p.268).

Conclusion

The justification of, and motivation for, criminal punishment, whether as a general practice or in specific application, may seem easy to dismiss as mere pontification from the ivory tower. But institutions, practices, rules and penalties in the criminal justice are chosen for a reason (or reasons), and

inquiring into those reasons reveals much about the particular criminal justice system itself. For instance, a system that widely employs plea bargaining is one that believes in making trade-offs amongst individual punishments, letting the small fish go now in hopes of catching a larger one later (Kipnis 1976); this acceptance of trade-offs is emblematic of a utilitarian/deterrence approach to punishment. On the other hand, a system that employs capital punishment despite the enormous material costs involved, and the questions regarding its deterrent value (Ehrlich and Liu 2006:v.III), displays a retributivist orientation that downplays the fact of material scarcity in favor of punishing based on desert.

Also, it must be realized that any system of criminal punishment does not operate in a vacuum, and it is often limiting to study and analyze it in isolation from the overall criminal justice, legal, and political systems. (This point is a central theme of Braithwaite and Pettit 1990.) As noted above, plea bargaining does not deal solely with punishment, but also prosecutorial discretion and the criminal trial process in general. Also, the question of appropriate punishment can be considered only after the relevant offenses have been defined as crimes, a problem that is at the heart of criminal law, which itself is a part of the larger political framework of society (as did both Bentham and Kant). In this sense, criminal punishment is linked inextricably to both moral and political philosophy, and no discussion of the topic can be truly complete without consideration of both.

Selected References

- Acton, H.B. (1969) (Editor) *The Philosophy of Punishment: A Collection of Papers*. New York: St. Martin's Press.
- Avio, K.L. (1990) "Retribution, Wealth Maximization, and Capital Punishment: A

- Law and Economics Approach", *Stetson Law Review*, Volume 19, 373-409.
- Avio, K.L. (1993) "Economic, Retributive and Contractarian Conceptions of Punishment", *Law and Philosophy*, Volume 12, 249-286.
- Beccaria, Cesare. (1764) *On Crimes and Punishments*, transl. David Young. Indianapolis: Hackett Publishing, 1986.
- Becker, Gary S. (1968) "Crime and Punishment: An Economic Approach", *Journal of Political Economy*, Volume 76, pp. 169-217.
- Bedau, Hugo Adam. (1978) "Retribution and the Theory of Punishment", *The Journal of Philosophy*, Volume 75, pp 601-620.
- Benn, S.I. (1958) "An Approach to the Problems of Punishment", *Philosophy*, Volume 33, pp. 325-341.
- Bentham, Jeremy. (1781) *The Principles of Morals and Legislation*. Buffalo: Prometheus Books, 1988.
- Bosanquet, Bernard. (1965) *The Philosophical Theory of the State*. London: Macmillan.
- Braitghwaite, John, and Pettit, Philip. (1990) *Not Just Deserts: A Republican Theory of Criminal Justice*. Oxford: Clarendon Press.
- Byrd, B. Sharon. (1989) "Kant's Theory of Punishment: Deterrence in Its Threat, Retribution in Its Execution", *Law and Philosophy*, Volume 8, Number 2, pp. 151-200.
- Cahill, Michael T. (2007) "Real Retributivism", *Washington University Law Review*, Volume 85, pp. 815-870.
- Card, Claudia. (1975) "Retributive Penal Liability", *American Philosophical Quarterly Monographs*, Number 7.
- Cooper, David E. (1971) "Hegel's Theory of Punishment", in Z.A. Pelczynski, ed., *Hegel's Political Philosophy: Problems and Perspectives*. Cambridge: Cambridge University Press.
- Cottingham, John. (1979) "Varieties of Retribution", *Philosophical Quarterly*, Volume 29, pp. 238-246.
- Davis, Lawrence H. (1972) "They Deserve to Suffer", *Analysis*, Volume 32, pp. 136-140.
- Davis, Michael. (1983) "How to Make the Punishment Fit the Crime", *Ethics*, Volume 93, pp. 726-752.
- Davis, Michael. (1986) "Harm and Retribution", reprinted in Simmons *et al* (1995), pp. 188-218.
- Deigh, John. (1984) "On the Right to Be Punished: Some Doubts", *Ethics*, Volume 94, pp. 191-211.
- Denning, Lord. (1984) *Report of the Royal Commission on Capital Punishment*. London: HMSO.
- Dolinko, David. (1991) "Some Thoughts About Retributivism", *Ethics*, Volume 101, pp. 537-559.
- Dolinko, David. (1997) "Retributivism, Consequentialism, and the Intrinsic Goodness of Punishment", *Law and Philosophy*, Volume 16, pp. 507-528.
- Duff, R.A. (1986) *Trials and Punishments*. Cambridge: Cambridge University Press.
- Duff, R.A. (2001) *Punishment, Communication, and Community*. Oxford: Oxford University Press.
- Duff, R.A. and D. Garland. (1994) (Editors) *A Reader on Punishment*. Oxford: Oxford University Press.
- Ehrlich, Isaac, and Zhiqiang Liu. (2006) (Editors) *The Economics of Crime*. 3 vols. Cheltenham, UK: Edward Elgar.
- Feinberg, Joel. (1965) "The Expressive Function of Punishment", reprinted in *Doing and Deserving*. Princeton, NJ: Princeton University Press, pp. 95-118.
- Finnis, J.M. (1972) "The Restoration of Retribution", *Analysis*, Volume 32, pp. 131-135.

- Flew, Antony. (1954) "The Justification of Punishment", reprinted in Acton (1969), pp. 83-102.
- Friedman, D. and W. Sjoström. (1993) "Hanged for a Sheep: The Economics of Marginal Deterrence", *Journal of Legal Studies*, Volume 12, pp. 345-66.
- Goldman, Alan H. (1979) "The Paradox of Punishment", reprinted in Simmons *et al* (1995), pp. 30-46.
- Hampton, Jean. (1984) "The Moral Education Theory of Punishment", reprinted in Simmons *et al* (1995), pp. 112-142.
- Hart, H.L.A. (1968) *Punishment and Responsibility: Essays in the Philosophy of Law*. Oxford: Oxford University Press.
- Hegel (1821) *The Philosophy of Right*, trans. T.M. Knox. Oxford: Oxford University Press, 1952.
- Hill, Jr., Thomas E. (1999) "Kant on Wrongdoing, Desert, and Punishment", *Law and Philosophy*, Volume 18, pp. 407-441.
- Honderich, T. (1984) *Punishment: The Supposed Justifications*. Revised Edition. Harmondsworth: Penguin Books.
- Kant, Immanuel. (1797) *The Metaphysics of Morals*, trans. Mary Gregor. Cambridge: Cambridge University Press, 1996.
- Kaplow, Louis, and Shavell Steven. (2002) *Fairness Versus Welfare*. Cambridge, MA: Harvard University Press.
- Kipnis, Kenneth. (1976) "Criminal Justice and the Negotiated Plea", *Ethics*, Volume 86, pp. 93-106.
- Kleinig, John. (1973) *Punishment and Desert*. The Hague: Martinus Nijhoff.
- Mabbott, J.D. (1939) "Punishment", in Acton (1969), pp. 39-54.
- Mackie, J.L. (1985) *Persons and Values*. Oxford: Clarendon Press.
- Moore, Michael. (1993) "Justifying Retributivism", *Israel Law Review*, Volume 27, pp. 15-49.
- Morris, Herbert. (1968), "Persons and Punishment", reprinted in *On Guilt on Innocence* (1976), Berkeley: University of California Press, pp. 31-58.
- Murphy, Jeffrie G. (1971) "Three Mistakes about Retributivism", *Analysis*, Volume 31, Number 5, pp. 166-9.
- Murphy, Jeffrie G. (1972) "Kant's Theory of Criminal Punishment", in Lewis White Beck (Editor), *Proceedings of the Third International Kant Congress*. D. Reidel, 434-441.
- Murphy, Jeffrie G. (1973) "Marxism and Punishment", in Simmons *et al* (1995), pp. 3-29.
- Murphy, Jeffrie G. (1987) "Does Kant Have a Theory of Punishment?" *Columbia Law Review*, Volume 87, pp. 509-532.
- Murphy, Jeffrie G. and Jean Hampton. (1989) *Forgiveness and Mercy*. New York: Cambridge University Press.
- Nozick, Robert. (1974) *Anarchy, State, and Utopia*. New York: Basic Books.
- Nozick, Robert. (1981) *Philosophical Explanations*. Cambridge MA: Belknap Press.
- Polinsky, A. Mitchell and Steven Shavell. (1979) "The Optimal Tradeoff Between the Probability and Magnitude of Fines", *American Economic Review*, Volume 69, pp. 880-891.
- Polinsky, A. Mitchell and Steven Shavell. (1999) "On the Disutility and Discounting of Imprisonment and the Theory of Deterrence", *Journal of Legal Studies*, Volume 28, Number 1, pp. 1-16.
- Posner, Richard A. (1980) "Retribution and Related Concepts of Punishment", *Journal of Legal Studies*, Volume 9, pp. 71-92.
- Rawls, John. (1955) "Two Concepts of Rules", *The Philosophical Review*, Volume 64, pp. 3-32.
- Rawls, John. (1964) "Legal Obligation and the Duty of Fair Play", in Sidney Hook

- (Editor), *Law and Philosophy*. New York: New York University Press, pp. 3-18.
- Ross, W.D. (1930) *The Right and the Good*. Indianapolis: Hackett Publishing.
- Scheffler, Samuel. (1988) (Editor) *Consequentialism and Its Critics*. Oxford: Oxford University Press.
- Scheid, Don E. (1983) "Kant's Retributivism", *Ethics*, Volume 93, pp. 262-282.
- Sher, George. (1987) *Desert*. Princeton: Princeton University Press.
- Simmons, A. John; Marshall Cohen; Joshua Cohen and Charles R. Beitz. (1995) (Editors) *Punishment: A Philosophy & Public Affairs Reader*. Princeton NJ: Princeton University Press.
- Smart, J.J.C., and Bernard Williams. (1973) *Utilitarianism: For and Against*. Cambridge: Cambridge University Press.
- Ten, C.L. (1987) *Crime, Guilt, and Punishment*. Oxford: Oxford University Press.
- Ten, C.L. (1990) "Positive Retributivism", in Ellen Frankel Paul, Fred D. Miller, Jr. and Jeffrey Paul (Editors), *Crime, Culpability, and Remedy*. Oxford: Basil Blackwell, pp. 194-208.
- Tunick, Mark. (1992) *Punishment: Theory and Practice*. Berkeley: University of California Press.
- Von Hirsch, Andrew. (1976) *Doing Justice: The Choice of Punishments*. New York: Hill and Wang.
- Von Hirsch, Andrew. (1978) "Proportionality and Desert: A Reply to Bedau", *The Journal of Philosophy*, Volume 75, pp. 622-624.
- Von Hirsch, Andrew. (1985) *Past or Future Crimes: Deservedness and Dangerousness in the Sentencing of Criminals*. New Brunswick NJ: Rutgers University Press.
- Wertheimer, Alan. (1975) "Should Punishment Fit the Crime?" *Social Theory and Practice*, Volume 3, pp. 403-423.
- Wertheimer, Alan. (1976) "Deterrence and Retribution", *Ethics*, Volume 86, Number 3, pp. 181-199.
- White, Mark.D. (2009) "Retributism in a World of Scarcity", in Mark D. White (Editor), *Theoretical Foundations of Law and Economics*. Cambridge: Cambridge University Press.
- Wittman, Donald. (1974) "Punishment as Retribution", *Theory and Decision*, Volume 4, pp. 209-237.

Mark D. White

Department of Political Science, Economics
and Philosophy, College of Staten Island,
City University of New York, USA
profmdwhite@hotmail.com

Community Health and Medicine

Peter Muennig

Introduction

Community health is a discipline that concerns itself with improving the health of communities via health promotion, enhanced opportunities for community participation, improved access to medical care, and expanded community infrastructure. Community health is multidisciplinary, with concerns that range from education policy to economics to enhanced medical services (Geiger 2002).

Community health adheres strongly to the idea that “health is a state of complete physical, mental and social well-being, and not merely the absence of disease or infirmity”, the definition used in the preamble of the World Health Organization’s constitution in 1948 (2005). The term *community medicine* is sometimes used to refer more narrowly to the provision of medical services within a community. Thus, community medicine is one dimension of community health.

Communities can be defined by a wide array of criteria ranging from a group of people with a similar interest or disposition (e.g., breast cancer survivors) to a geographic area (e.g., neighborhood or health district). Most often, community health focuses on a geographically defined community that has undergone a needs assessment using a combination of qualitative and quantitative data. Because community medicine is more focused on local service delivery and the biomedical aspects of health, it nearly always addresses the needs of geographically defined communities.

Because the broader field of community health emphasizes non-medical aspects of health, national and local governance issues are seen as key ingredients in building

healthy communities. Healthy communities contain opportunities for meaningful participation in politics, provide residents with a sense of agency, tend to have strong social networks between individuals, have organizations that provide opportunities for social interaction, minimize discriminatory practices, and offer high quality public services, such as schools and health care institutions (Krieger & Sidney 1996; Kawachi, Kennedy et al. 1999; Geiger 2002; Jia, Muennig et al. 2004).

While the goals of community health practitioners are similar, the endpoints sometimes differ by discipline. Urban policy researchers and economists sometimes see the intimate link between social well-being and disease as just one of many benefits to improving a community’s infrastructure. Public health practitioners, on the other hand, view improved health (measured as changes in morbidity or mortality rates) as the outcome of interest.

Regardless of the community health practitioner’s discipline, however, the assumption behind community health is that health arises when people optimize environmental conditions—an assumption that is intuitive and supported both by research studies and inferences drawn from history (Wilkinson 1999; Geiger 2002).

History

Modern civilization first arose from communities—groups of humans that worked together to hunt, gather, and prepare food as well as fend off predators. (See Diamond 1998 for an excellent review.) Collective organization allowed a transition from hunting and gathering to farming and food storage. Farming in turn allows people to remain in one location, thus allowing for the construction of dwellings that protect people from the elements. Finally, collective organization afforded the use of increasingly

complex tools that could be continuously improved with the input of other community members. Production thus became increasingly efficient, freeing up time for the production of goods for trade. Trade in turn improves the quality of textiles, diversifies the diet, and improves the exchange of technical information between communities, providing more subtle improvements in the quality and quantity of life.

Soon, though, humans became victims of their own success. For instance, increased population density probably enhanced the development and spread of infectious disease. Collective living also created problems with the acquisition of clean water and disposal of feces. Moreover, communities allowed for the development of war and social hierarchies. The existence of social hierarchies in turn meant that only some community members were able to fully enjoy the benefits afforded by collective living. For some people (e.g., human slaves), the quality of life greatly deteriorated within the societies to which communities ultimately gave rise.

Each of these problems has since been tackled in novel ways. In the eighteenth and nineteenth centuries (primarily in Europe, Oceania, and North America), improvements in the quality and quantity of life tended to come in the form of advances in governance (such as the introduction of democracy, taxation, and public education) and technical advances (such as sewage).

Rudolf Virchow was one early observer of the impact of such phenomena on human health (Virchow 1985). As a young physician, he was dispatched to investigate an infectious disease outbreak in a community in Upper Silesia (which is now in Poland). He noted that the disease spread almost exclusively among the poor and less educated persons living in overcrowded conditions. He speculated that these problems arose from hierarchical inequities within the community

and recommended a wide array of interventions ranging from progressive taxation to the institution of free education and democracy within Germany.

Virchow was also an advocate for basic public health interventions, and helped institute and design Berlin's sewage system. Recognition of the benefits of such interventions in the late 1800s led to the widespread adoption of clean water, sanitation, and occupational safety regulations in industrialized nations. These measures may have been responsible for the sharp increase in life expectancy in industrialized nations over the decades spanning the late 19th and early 20th century, with some localities fully doubling life expectancy in the span of 20 years.

Pre-World War II progressive era policies in the United States, and post-war policies in Europe resulted in large increases in life expectancy. These included obligatory education, social security, and other welfare programs (which, in most nations, included universal healthcare). Collectively, these programs helped protect low-income communities from hunger while provided shelter and other necessities. These social programs were greatly helped along by the advent of vaccines and antibiotics.

Many of these public health, social, and medical advances would not have been possible without democracy and universal education (Sen 1993). Education not only broadens the pool of knowledge resources from which society can draw upon, but also provides the individual with cognitive skills that work to improve chances of survival. By giving the poor a voice, democracy helps ensure that social resources are made available to those that need them the most.

In poor communities within developing nations, improvements on the social, public health, and medical fronts came much more slowly. Within colonies of wealthy nations, it

was the colonists who enjoyed the privilege of democracy, housing, and social infrastructure. After World War II, the abrupt withdrawal of colonial powers left little in the way of human capital and knowledge. As a result, what little infrastructure that remained deteriorated in poor governance.

However, even some industrialized democracies failed to fully develop the redistributive mechanisms needed to give poor communities full opportunities for growth and development. For instance, the chance of a black male reaching age 65 in Harlem, New York is lower than that for males in Bangladesh (McCord and Freeman 1990). Within these communities, the percent immunized, voting rates, and school quality parallel similar indicators in communities within developing countries (Murray, Lopez et al. 1996). Why participation by low-income communities in the democratic process remains low even after the implementation of universal suffrage is an area that requires more research (Verba et al 1995).

Because there is sometimes less will on the part of governments to invest in healthy communities within poorer neighborhoods, community health interventions in the latter half of the 20th century were often forwarded by individuals or groups rather than governments. In the 1950s, Sidney and Emily Kark developed a model for integrating clinical medicine and public health activities similar to those proposed by Virchow nearly one hundred years before. Called community-oriented primary care, this model of health care delivery included assessment of the health needs of a community and prioritization of health projects based on this assessment. While this model focuses more on medical care, improvement of the socio-economic status of community members was implicitly recognized as central to improving the health of the community as a whole

(Tollman et al. 1997). Their project in South Africa was shut down by the Apartheid government, but has been applied with great success elsewhere.

In the late 1950s and 1960s, a host of new models for community well-being emerged. For instance, Paulo Freire's idea that persons in impoverished communities can achieve agency through critical dialog opened the doors to building human and social capital within resource poor communities (Daniels 1969). These models were put to practice in Bolivar County in the state of Mississippi in the United States (Geiger 2002). There, Jack Geiger and others created a broad health intervention based on the needs of the African-American population it served.

This intervention included the delivery of health services combined with community outreach, the implementation of a public transportation system, water pump installation, outhouse construction, and housing improvements among many other social interventions. This participatory model was focused on empowerment and included the development of a food cooperative and buy-in from a large bank. This bank was offered the incentive of holding deposits on the community intervention in exchange for opening branch offices and providing loans at rates similar to those provided to higher income white communities. The combination of jobs created by the cooperative and capital provided by the bank allowed community residents to participate in the larger economy from which they had been removed. The intervention was an unprecedented success; not only did health indicators improve dramatically, but community members went off to study in college and to obtain advanced academic degrees. Some students returned with medical degrees.

The Bolivar Country project went beyond the traditional scope of community-oriented primary care, reinforcing the idea that

prevention is needed in conjunction with treatment. Eventually, the non-medical (often referred to as “social”) determinants of health came to be accepted within the discipline of public health (Strelnick 1999). Likewise, other disciplines began to recognize that social interventions also resulted in unexpected gains in health. As a result, older fields of study, such as social medicine, were revived and new disciplines, such as social epidemiology, were created (Krieger 2001).

At the turn into the 21st century, modern community medicine and public health increasingly incorporated novel technologies including molecular markers of disease and geographic information systems. Molecular disease markers allow for tracking of organisms’ genetic resistance to antimicrobials as well as tracking human-to-human spread. Geographic information systems have greatly facilitated our understanding of these data by allowing disease transmission or prevalence to be tracked between communities ranging from remote villages in Africa to neighborhoods within major urban centers, such as New York City. These systems are also being used to understand how the social or structural characteristics of communities influence health.

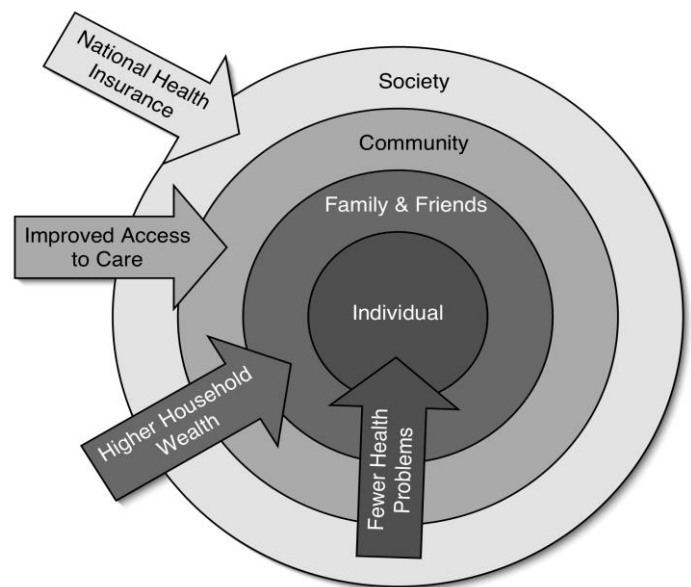
Human Capital and Community Health

As evidenced by the history of community health, most of the problems associated with community wellbeing arise from social problems inherent to the society within which the community resides. Issues surrounding governance at the societal level are therefore at least as critical as they are at the community level.

One theoretical model for conceptualizing the influence of societal factors on the community is presented in Figure 1, below. Here, we see the influence of society’s political economy on the health of

communities. Societal and community factors in turn influence the strength and scope of networks of family and friends. Ultimately, all these factors work together as determinants of the health and wellbeing of the individual. Thus, this figure can be thought of as a concentric Venn diagram, with the society exerting influence on the health determinants of communities, social networks, and individuals.

Figure 1. Influence of Various Forms of Social Structure on Community and Individual Health.



In this example, national health insurance leads to improved access to health care in the community. This free care in turn reduces total household expenses, producing higher household wealth. Finally, higher household wealth allows people to access to material goods that further improve health.

In *Figure 1*, we see that national health insurance has multiple effects on the socio-economic wellbeing of a community. Each arrow points to various rippling effects of national health insurance at the societal, community, family, and individual levels. For example, because illness has a disproportionate effect on the total wealth of the poor, the provision of universal health insurance serves as a redistributive mechanism by which low-income communities are made healthier and wealthier (see the second concentric circle). Of course,

insurance also exerts a direct effect on health via the provision of medical care (center circle). This in turn produces more productive individuals while reducing the economic burden of insurance on private enterprise. In sum universal health insurance exerts direct effects via access to needed medical care and indirect effects on health via improved earnings (through increased productivity) and improved wealth (through reduced health expenditures).

Of these societal and community factors responsible for improving health via human capital, few are as important as education (Muennig and Fahs 2001). While health enables people to participate in civic activities and the labor market, this participation also appears to be a key determinant of health (Kawachi et al. 1999).

Thus, one of the key effects of societal policies on community health is to build human capital. This is almost certainly best accomplished via guaranteed primary and secondary education (Schultz 1961). Nations that offer obligatory high quality education to all persons within their borders have the lowest rates of morbidity and mortality (Preston 1976). This is probably due in part to micro-economic factors, such as empowering individuals to participate in a meaningful way in the economy. But it is also likely due to macro-economic factors, such as building a robust economy in the first place.

A robust economy provides communities with access to nutritious food, quality health care, safe housing, safe transportation, and so forth. In this way, economically productive members of economically developed countries have achieved the “ecological niche” within which humans can optimize their chances of surviving (Wilkinson 1999). However, the health of a nation is dependent on the extent to which ecological factors are optimized across all communities.

Among very poor countries, per capita gross domestic product is a strong predictor of life expectancy. In this context, very few communities within a nation are prosperous enough to provide the basic ingredients for survival: safe food, clean water, sanitation disposal, and protection from disease vectors. Once a country has reached a point of relative development, it is the extent to which educational and health resources—and thus human capital—are distributed across communities that appears to matter most. In sum, human capital builds healthy communities via economic development, but it must be fairly distributed to maximize the health of all communities.

Much of this theory is speculative because it is not possible to draw concrete conclusions from cross-national studies. Nonetheless, nations such as Sweden, which allow all children roughly equal access to an education, do best in terms of life expectancy and infant mortality. Nations with more of an income-based system, such as the United States, fare far worse on these measures. (Though education is offered to all US citizens, its quality varies widely, with two students to a desk in some low-income communities. Moreover, other programs that build human capital, such as universal health insurance, are lacking.) At the bottom of the hierarchy are countries that have made very small investments in education historically, such as Sierra Leone or Liberia. In these instances, only the wealthiest communities have access to adequate schooling or other resources needed to build human capital.

One effect of human capital is that it allows community members to participate in their community in a vibrant way. Participation includes empowerment through an earned income, an equal vote, and a sense that the individual has adequate standing relative to other community members. These factors are part of what makes up a much

more amorphous concept termed *social capital*.

Social Capital and Community Health

Social capital refers to the social ties that bind communities (and, by extrapolation, societies) together. Social capital includes abstract human qualities, such as trust, confidence, caring, and advocacy. The loss of social capital is mourned by those to the left as well as to the right of the political spectrum. Those of the right attribute the decline to the loss of religion and/or family values and those on the left attribute it to the diminishing checks on market forces (Persell 1997). Those on the right argue that mass exposure to sexually explicit or otherwise immoral materials in the media results in the slow deterioration of values that promote social cohesion. Underpinning the argument on the left is an increasing emphasis on individualistic survival in a competitive marketplace, which is driving up work hours and thereby reducing time for friends, family, and community activities.

Regardless of the cause, communities that lack access to social resources designed to build human capital also tend to have little in the way of social capital. To the extent that such communities are economically disenfranchised, it is logical that they suffer from crime, poor nutrition, and other threats to health. Less clear, though, are links between other forms of social capital and health. Strong social networks, perceived trust, and optimism all appear to be strongly associated with lower mortality. The excess mortality resulting from a breakdown in this form of social capital mostly assumes the form of heart disease, cancer, and diabetes (Wong et al 2002). While partly explained by environmental stressors and risky health behaviors, the causes of this link remain an enigma.

Many concrete examples of the inter-relationship between human capital, social capital, and health exist. For instance, in most industrialized nations, it would have been unthinkable that a government could deny the existence of HIV/AIDS within its borders precisely because there is sufficient knowledge of the properties of the disease (human capital) that it would be difficult to dupe the citizenry *en mass*. Likewise, social capital (in the form of emotional support) might hinder markets for risky practices, such as prostitution or intravenous drug use. In nations where only a minority of the population is well educated and social ties have been broken down, disinformation campaigns probably helped HIV/AIDS gain a foothold, resulting in a good deal of suffering and economic damage.

The study of the relationship between social capital and health has a number of serious flaws, however. First, there may be important differences between concepts like strong social networks (which exist both due to family factors and community factors) and social trust. These concepts are based on self-report data, and are typically analyzed using regional comparisons (McKenzie et al 2002).

Nonetheless, the study of social capital is central to understanding what makes communities healthy; it is the collective values of individuals within a community that determines their willingness to advocate for those very things that make a community healthy. Moreover, the strong and consistent association between perceptual factors or social network size and health—which exist both across communities and in prospective scientific experiments—cannot be ignored (Cohen et al 2003).

International declines in social capital have been observed over the past few decades, with the epidemic being driven by low-income communities within countries with poor social infrastructure and the United

States. Crime rates have exploded from Lagos' giant shantytown corridors to the slums of Mexico City (Davis 2004). Kidnappings for ransom have become a part of daily life in numerous countries. In the United States, violence has been contained by aggressive policing and imprisonment. However, trust and happiness have continued their downward slope (Putnam 1995; Kawachi et al 1999).

In cross-national studies, income inequality has been linked to crime rates (Hsieh and Pugh 1993). Here it is hypothesized that unjust exclusion from participation in the economy has led to a collective anger, causing a breakdown in community norms. Supporting this notion is the observation that robust civil society exists in countries that have invested heavily in social programs. However, the existence of a robust civil society in countries with social programs that reduce income inequality may be as much a reflection of the norms that created the programs as the programs' influence on social norms in the first place.

Racial and ethnic cohesion is an important component of the social capital that glues communities together. Racial identity can be used to build cohesion via shared culture, art, and heritage. But it can also be used to exclude persons of other races from accessing educational or economic activities, resulting in a breakdown of social capital, and thus social order. In the United States, for instance, blacks have historically been exposed to discriminatory housing and employment policies that have created ghettos with inadequate schools. It is perhaps because of the exclusionary policies that homicide is the leading cause of death among 15 to 35 year-old African-American males (Arias et al 2003).

While violent crime poses a growing threat to human life, people in high crime neighborhoods suffer disproportionately from

disease rather than injury. A portion of these deaths may be due to the lack of availability of healthy foods within poor neighborhoods. Another portion is likely due to the lack of opportunities for exercise (due in part to high crime rates) and to high rates of smoking (Lantz et al 1998).

Pessimism, skepticism, and other "negative emotions", are more prevalent in low-income communities and are strong predictors of mortality (Cohen et al 2003). Those with strong and large social networks tend to be at significantly lower risk of heart disease than those who do not (Berkman 1982). Likewise, stress associated with living in poor neighborhoods has been proposed as a major determinant of morbidity and mortality. These factors are all logically linked together; healthy communities with a wide array of opportunities for social interaction likely increase trust, foster optimism, build friends—all things that reduce innate stress and provide cushions for stressful life events. Conversely, high crime neighborhoods with few friendly institutions foster loneliness (an independent risk factor for heart disease) and isolation in addition to providing a wide array of novel stressors (Taylor 2002). Stressful life events include the lack of available childcare, lacking control in the job, and the inability to pay the bills at home. Crime, poor housing stock, pest infestation, and noise may also contribute to the high levels of self-rated stress among residents of low-income communities.

If declines in social capital are occurring, and these in turn are hindering improvements in life expectancy, the relevant question is whether these trends can be reversed, and how. A wide array of potentially causative agents have been proposed, including removing violence from television, reducing the media's inclination to scare the public into tuning in to their programming, imposing ratings on compact disks, and so forth.

However, the correlation between neighborhood income, neighborhood social capital, and poor health outcomes suggests that the solution might lie in redistributive policy.

Models and their Effectiveness

Models for community development, including those forwarded by community-oriented primary care (community-oriented primary care--see History above), demonstrate that social capital can be rebuilt, with concomitant improvements in community health and wellbeing. Remedies for ill communities range from small-scale local interventions to large-scale nation-wide programs, in which many different government services are geared toward building healthy populations at the community level (Acheson 2002).

While there are many examples in which healthy communities have risen from the ashes of crime and sickness, there is no single prescriptive model that will work in all instances. Communities vary greatly in terms of needs, demographics, and capacities. Standardized remedies are thus less likely to work, even when applied to communities within a single nation.

Moreover, while communities that have received public health, urban development, or community-oriented primary care based interventions have been observed to improve, it is nearly always impossible to prove beyond doubt that the intervention was the causative factor.

Evaluation is difficult when discussing smaller scale interventions because there are many competing factors within a community that influence health outcomes. Examples of small-scale interventions include training community members to become community health workers who work to improve health knowledge within a community, or patient navigators who walk community members

through the process of medical screening. In each case, outcomes such as increases in health knowledge or the number of tests administered are difficult to compare with competing health needs.

One large-scale program, the Neighborhood Strategy for Social Renewal in the UK, brings community members, government organizations, and community-based organizations to the table to address education, crime, housing, and medical needs. This program is run jointly by health and community development authorities, and offers US\$90 million to agencies within each targeted community. This program monitors multiple levels of outcomes to determine efficacy, with measures of inter-agency cooperation at the beginning of the assessment chain and health outcomes at the end. Thus, success at interagency coordination is linked to intermediate outcomes (e.g. graduation rates) that are in turn linked to health outcomes. Should pathways receiving interventions unfold in a clear and logical way from distal to proximal outcomes, it is easier to qualitatively infer success.

Typically, multifaceted or broader scope interventions are needed to produce visible results that are more easily linked to the intervention, even when they are relatively small in scope. The first community-oriented primary care intervention in South Africa focused on assessing and addressing the health needs of the community, including viable education services. The intervention appears to have resulted in longstanding improvements in wellbeing, even though the Apartheid government quickly shut down the program (Tollman et al 1997). Likewise, preschool intervention programs have shown lifelong improvements in criminal behavior and employment, albeit modest (Karoly & Bigelow 2005).

While some argue that other factors could account for the success of many of these programs (e.g., chance neighborhood gentrification or long-term trends toward economic improvement), some policymakers have found the models convincing enough to merit large public outlays.

Building on the concept of community-oriented primary care, the Labor government of the UK commissioned a comprehensive study of the impact of inequalities in health (Acheson 2002). The report recommended that virtually all government sectors consider the potential health impact of policies thought to influence inequality, and that coordination between agencies explicitly for this goal be implemented. Moreover, the Acheson report broadly defines communities, explicitly including ethnic groups, women, and the elderly.

Some of the recommendations of the Acheson report have since been implemented (Marmot 2004). Government programs, such as the Sure Start preschool program for poor children, are now to be designed with consideration of their impact on the health of communities they serve. Essentially using a community-oriented primary care model, Sure Start not only enhances education, but combines them with other services needed for the education to become effective. For instance, preschool programs have to be combined with daycare for younger children if parents are to participate in the workforce, thus improving the family's financial position. Most importantly, Sure Start is community-based and works using a participatory framework (parents provide feedback on the direction and needs of the program). Mandates similar to the Acheson report have been passed in Europe, Canada, and Australia.

At the community level, various renditions of the community-oriented primary care model have been implemented in a wide array

of settings. Most experience with this model is derived from the developing setting, where many non-governmental organizations utilize a participatory framework before acting. This framework involves obtaining the input of community members so that needs can be assessed. Perhaps more importantly, it ensures "buy in" by giving community members a role in making the changes for which they have advocated. "Buy in" is a means of giving employees the sense of a personal investment in the outcomes of the program, thus increasing employee enthusiasm and input.

For instance, Tendler examined a government intervention in Brazil in which community members were trained as community health workers to increase vaccination rates (Tendler 1998). Because these programs typically involve money funneled from the central government, the process can become political, bureaucratic, and open to cash siphoning or inappropriate use of the facilities provided. In this instance, a local political leader built the intervention around the outcome (increased vaccination rates) and made the community health worker position open and competitive (rather than nepotistic). This gave the activity a sense of importance, and people came together under the goal of preventing disease. Word of the importance of the project spread from the community health workers (who lived in the community and knew community members), so parents of the children to be vaccinated bought in as well. By the end of the campaign, vaccination rates had gone from among the lowest to among the highest in the country.

Because community health entails interventions directed at neighborhoods rather than people, hard experimental evidence of the effectiveness of community health interventions is difficult to come by. The weakest study design simply follows changes

in a community after implementation of the program. Since the health of communities generally improves over time, it is difficult to determine which effects were due to the intervention and which were not. (In extreme instances, such as the Bolivar County example above, it's nonetheless quite obvious.)

When community interventions are formally studied, the experimental community is sometimes compared to control communities. While better than monitoring changes in one community over time, a wide array of ecological factors can complicate the comparisons. Some of these effects can be controlled for in statistical models, but some cannot be foreseen.

Individual level experiments, such as the randomized controlled Moving to Opportunity (MTO) program, evaluate the effects of programs designed to improve the health of individuals via improvements in their environment (Kling et al 2004). The MTO program evaluated the effect of randomly providing housing vouchers to residents of low-income neighborhoods in various cities in the United States versus assigning them to public housing. Thus, subjects were allowed to choose private housing outside of their dangerous communities. While hardly helping the community from which the residents move, this program demonstrated that the health of individuals within communities can improve with improvements in housing.

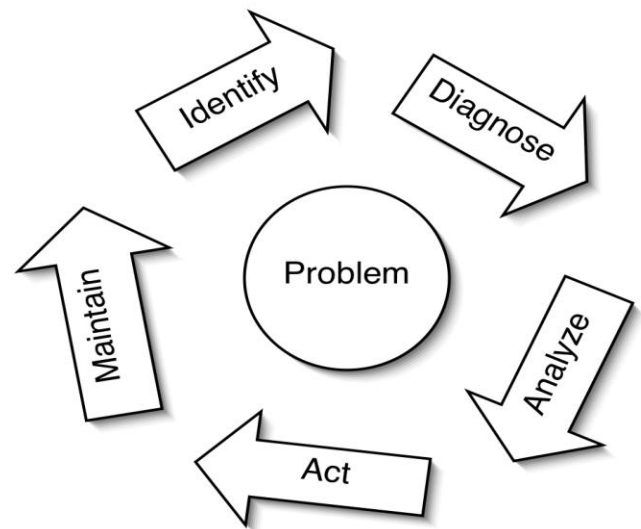
There is also evidence that income supplementation programs are effective at improving some aspects of the health of individuals within communities (Kehrer & Wolin 1979). Such interventions are a far cry from the multifaceted interventions that build individual agency inherent to community-oriented primary care programs, so their efficacy is especially revealing. Likewise, in a non-randomized study, Costello et al (2003)

found evidence linking income supplementation to improvements in social pathology (Costello et al 2003).

Practice

Community-oriented primary care, urban planning, and other disciplines all use a similar model for planning and implementing community interventions. See Figure 2, below.

Figure 2. Community Action Model.



In this model, the problem is identified, data are collected, data are analyzed, the program is implemented, and the program is maintained and re-evaluated.

This begins with assessing a given community to ascertain how to optimize conditions (Kark 1974). Second, community “diagnosis” is conducted in which the roots of the identified problem(s) are examined. Third, qualitative and quantitative analyses are sometimes conducted to ensure that the problem is indeed real, to fully understand all dimensions of the problem, and to ascertain the extent of the problem. Fourth, actions (also called interventions or strategies) are selected to address the problem, ideally with the input and participation of community members. Fifth, the action is maintained and periodically assessed to ensure its effectiveness. Examine Figure 2 again. The

process is then repeated based on the results of periodic assessments. This model, sometimes called the community action model, is more succinctly stated as “look, think, act, reflect” (Koch & Kralik 2001).

There are various tools used to identify problems in the community that are directly or indirectly related to health (the first step in *Figure 2*). On the technical side, these include geographic information systems that map community statistics and provide comparisons between communities. Inputs include census data, medical billing data, and communicable disease surveillance data, among others. Community-based organizations (or non-governmental organizations), academic institutions, the media, law enforcement agencies, and government representatives also collect quantitative and qualitative data to help identify problems within communities.

Diagnosing problems (the first step in the community action model) begins with conceptualizing community interventions in terms of their scope (i.e., multifaceted or one-dimensional) and framework (e.g., participatory). Communities with one overriding simple problem are rare. However, most interventions that address multiple aspects of a given problem (e.g., the coupling of education, jobs programs, daycare, and housing) may prove to be more cost-effective than single interventions alone.

For instance, in the economic south (developing nations), building schools in impoverished communities does little when the children are too hungry and too sick to attend school. In industrialized nations, jobs programs do little good when the parent has children at home. In other words, combining two effective programs often produces greater benefit than the total benefit of each program alone. An often-cited example is society’s focus on building medical clinics at the expense of environmental interventions (e.g.,

mosquito eradication) that would prevent the very conditions the clinics treat. Nevertheless, organizations that do one thing and do it well may be more effective if they collaborate with other organizations rather than expanding into unknown territory. For instance, medicine-based AIDS treatment programs will benefit from collaboration with organizations working to improve nutrition.

A second component of the assessment and planning phase is deciding upon a framework within which to act. The Tendler case above highlights the utility of community participation. Participatory rural appraisal is a key component of community health. It is an all-encompassing term covering a broad array of techniques for getting community input and involvement for health interventions (Chambers 1994). While participatory rural appraisal ideally involves empowering community members to design and implement the interventions they benefit from, it more often occurs with the input or even direction by community outsiders, who control cash flows. Nonetheless, drawing upon local knowledge and the community workforce reduces the likelihood of unforeseen glitches and usually greatly reduces personnel costs by relying upon local salaries rather than those of industrialized nations. The Bolivar County (in the State of Mississippi in the United States) intervention described above used a similar model in the 1960s, and many earlier instances no doubt exist. However, participatory rural appraisal came into vogue in the late 1980s, perhaps in part due to the World Health Organization’s Declaration of Alma-Ata, which recommends a participatory framework in primary care delivery (WHO 1978).

But participatory rural appraisal is not often practiced, and sometimes not practical. Community health practitioners often have an area of interest (sometimes a mandate) that necessitates matching communities to the

practitioner's interest, rather than allowing the community to define its needs. For instance, food aid programs necessarily deal with nutrition, whether or not it is actually a priority in a given community. Other times, monitoring agencies, such as local health departments, identify immediate threats, such as infectious disease outbreaks, and address them as quickly as possible. (While it would be ideal to have participatory community-based infrastructure ready to address such problems, this is often not practical, given budgetary constraints.)

Another example is an intervention within immigrant communities that requires the use of outsiders familiar with the larger cultural context. Such "cultural navigators", by definition, cannot be hired from the communities targeted by the intervention.

There are also many examples of how community members can participate in the intervention with outside control, but community participation. One example is a patient navigator intervention. In this instance, a community health worker is trained to perform a specific task, such as teaching parishioners in a church how to access preventive medical care. The community health worker is drawn from the church and is thus familiar with the community, but the intervention can be directed by the medical clinic offering the preventive services.

The action component of the community action model involves implementation of the hard work done in assessment of the needs of communities and planning the intervention. Whatever the action is, it should be coupled with an evaluative component to make sure that it is working. Qualitative data are essential in refining the model at this point; the planners will want to know what is working well, what should be dropped, and what might be changed or added. With this, the cycle begins again.

Conclusion

Given that groups with common culture or interests tend to live close to one another, community health interventions can sometimes be neatly targeted to a particular geographic region. More often, communities are too diverse for simple strategies. For instance, within Chinatown in New York City, multiple Chinese languages are spoken, rendering culture-specific or radio-based interventions effective only for sub-populations. These issues are especially prescient in ethnically or religiously diverse countries, exemplified by Canada, the United States, or England among industrialized nations and India and Indonesia among developing nations.

Communication amongst geographically dispersed communities has been greatly facilitated by the Internet (typically accessed via Internet cafés in the developing world). The Internet has provided a hub for information exchange only previously available through organizations that relied on mail or phone communications. This has allowed communities from different regions to share information, to come together for participation in online information exchange, and to quickly mobilize for social movements.

Regardless of the type of community being targeted for a health intervention, it is critical to avoid doing more harm than good. Countless examples of well-intended interventions gone wrong (some unavoidable and some careless) abound in the development literature. In one instance, water pumps were provided to communities residing above arsenic-laden groundwater. In another, a tsetse fly eradication program allowed communities to relocate to the sides of rivers only to contaminate drinking water for downstream communities. Finally, countless economic initiatives have been

undertaken in the name of “poverty reduction” based on arbitrary standards (e.g., communities with mean earnings of less than \$1 per day). In many such instances, the community members considered themselves to be well off until the programs were implemented, removing them from their sustenance lifestyle to work for sometimes ruthless employers in the marketplace.

Most of these pitfalls can be avoided with careful planning, the use of a participatory framework, and detailed community assessments that include qualitative data in addition to quantitative data. Common sense dictates avoiding repeating the mistakes that have already been made by industrialized nations: a switch from healthy forms of transport to car-based forms, employing welfare policies that emphasize dependency rather than autonomy, and failing to consider multiple dimensions of a given social problem (e.g., failing to couple earned income tax credits with the daycare, education, public transportation, and healthcare needed to make it possible for people to work in the first place). Multifaceted, well-designed community interventions that employ a participatory framework have in many instances proven beyond reasonable doubt to dramatically increase the length and quality of the lives of the people that make up communities worldwide.

Selected References

- Acheson, Donald. (2002) *Independent Inquiry into Inequalities in Health Report*. London: The Stationary Office, 164ff.
- Arias, Elizabeth and Robert Anderson. (2003) “Deaths: Final Data For 2001”, *National Vital Statistics Reports*, 52, 3, 116ff.
- Berkman, Lisa. (1982) “Social Network Analysis and Coronary Heart Disease”, *Advances In Cardiology*, 29, 37-49.
- Chambers, Robert. (1994) “The Origins and Practice of Participatory Rural Appraisal”, *World Development*, 22.7, 953-69.
- Cohen, Sheldon; William Doyle; Ronald Turner; Cuneyt Alper; and David Skoner. (2003) “Sociability and Susceptibility to the Common Cold”, *Psychological Science*, 14, 5, 389-95.
- Cohen, Sheldon; William Doyle; Ronald B. Turner; Cuneyt Alper; and David Skoner. (2003) “Emotional Style and Susceptibility to the Common Cold”, *Psychosomatic Medicine*, 65, 4, 652-7.
- Costello, Jane E.; Scott N. Compton; Gordon Keeler and Adrian Angold. (2003) “Relationships between Poverty and Psychopathology: A Natural Experiment”, *JAMA*, 290, 15, 2023-9.
- Daniels, Robert. (1969) “Toward a New Model of Human Service Delivery”, *Health Services Research*, 4, 2, 91-5.
- Davis, Mike. (2004) “Planet of Slums. Urban Involution and the Informal Proletariat”, *New Left Review*, 26, 5-34.
- Diamond, Jared. (1998) *Guns, Germs, and Steel*. London, Random House.
- Geiger, H. Jack. (2002) “Community-Oriented Primary Care: A Path To Community Development”, *American Journal Of Public Health*, 92, 11, 1713-6.
- Hsieh, Ching Chi. and M. Pugh (1993) “Poverty, Income Inequality, and Violent Crime: A Meta-Analysis of Recent Aggregate Data Studies”, *Criminal Justice Review*, 18, 182-202.
- Jia, Haomiao; Peter Muennig; E. Lubetkin and M. Gold (2004) “Predicting Geographical Variations in Behavioural Risk Factors: An Analysis of Physical and Mental Healthy Days”, *J Epidemiology And Community Health*, 58, 2, 150-5.
- Kark, Sydney. (1974) *Epidemiology and Community Medicine*. New York, Appleton-Century-Crofts.

- Karoly, Lynn and James E. Bigelow. (2005) *The Economics of Investing in Universal Preschool Education in California. Rand Labor And Population*. Santa Monica, Rand Corporation, 238ff.
- Kawachi, Ichiro; Bruce P. Kennedy and R. Glass. (1999) "Social Capital and Self-Rated Health: A Contextual Analysis", *American Journal of Public Health*, 89, 8, 1187-93.
- Kawachi, Ichiro and Bruce. P. Kennedy. (1999) "Crime: Social Disorganization and Relative Deprivation", *Social Science And Medicine*, 48, 6, 719-31.
- Kehrer, Barbara, and Charles. M. Wolin (1979) "Impact of Income Maintenance On Low Birth Weight: Evidence From the Gary Experiment", *Journal of Human Resources*, 14, 4, 434-62.
- Kling, Jeffrey; Jeffrey Liebman; Lawrence F. Katz and Lisa Sanbonmatsu. (2004) "Moving to Opportunity and Tranquility: Neighborhood Effects on Adult Economic Self-Sufficiency and Health From a Randomized Housing Voucher Experiment". *NBER, KSG Working Paper* 57.
- Koch, Tina, and Debbie Kralik (2001) "Chronic Illness: Reflections on a Community-Based Action Research Programme", *Journal of Advanced Nursing*, 36, 1, 23-31.
- Krieger, Nancy (2001) "Theories for Social Epidemiology in the 21st Century: An Ecosocial Perspective", *International Journal of Epidemiology*, 30, 4, 668-77.
- Krieger, Nancy, and Stephen Sidney (1996) "Racial Discrimination and Blood Pressure: The CARDIA Study of Young Black and White Adults", *American Journal of Public Health*, 86, 10, 1370-8.
- Lantz, Paula; Jeffrey S. House; J.M. Lepkowski; D.R. Williams; R.P. Mero and Chen J. (1998) "Socioeconomic Factors, Health Behaviors, and Mortality: Results From a Nationally Representative Prospective Study of US Adults", *JAMA*, 279, 21, 1703-8.
- Marmot, Michael. (2004) "Tackling Health Inequalities Since the Acheson Inquiry", *J Epidemiology And Community Health*, 58, 4, 262-3.
- Mccord, Colin, and Harold P. Freeman. (1990) "Excess Mortality in Harlem", *New England Journal of Medicine*, 322, 3, 173-7.
- Mckenzie, Kwame; Rob Whitley and Scott Weich. (2002) "Social Capital and Mental Health", *British Journal Of Psychiatry*, 181, 280-3.
- Muennig, Peter, and Marianne Fahs (2001) "The Cost-Effectiveness of Public Postsecondary Education Subsidies", *Preventive Medicine*, 32, 2, 156-62.
- Murray, Christopher and Alan Lopez; (1996) *The Global Burden of Disease: A Comprehensive Assessment of Mortality and Disability From Diseases, Injuries, and Risk Factors in 1990 and Projected to 2020*. Cambridge, MA, Harvard School of Public Health on Behalf of the World Health Organization and the World Bank; Harvard University Press.
- World Health Organization (2005) *About WHO*. 2005. www.who.int/about/en/
- Persell, Caroline H. (1997) "The Interdependence of Social Justice and Civil Society", *Sociological Forum*, 12, 2, 149-72.
- Preston, Samuel H. (1976) *Mortality Patterns in National Populations: With Special Reference to Recorded Causes of Death*. New York, Academic Press.
- Putnam, Robert (1995) "Tuning In, Tuning Out: The Strange Disappearance Of Social Capital in America", *PS, Political Science & Politics*, 28, 664-78.
- Schultz, Theodore (1961) "Investment in Human Capital", *American Economic Review*, 1-17.

- Sen, Amartya (1993) "The Economics of Life and Death", *Scientific American*, 268, 5, 40-7.
- Strelnick, Hal (1999) "Community-Oriented Primary Care. The State of an Art", *Archives Of Family Medicine*, 8, 6, 550-2.
- Taylor, Humphrey (2002) "Poor People and African-Americans Suffer the Most Stress From the Hassles of Daily Living". *The Harris Poll* 66. Rochester, NY, Harris Interactive.
- Tendler, Judith (1998) *Good Government in the Tropics*. Baltimore, Johns Hopkins University Press.
- Tollman, Stephan, Kark, Sydney, Et Al. (1997) *The Pholela Health Centre: Understanding Health and Disease in South Africa Through Community-Oriented Primary Care (COPC)* Oxford, UK, Clarendon Press.
- Verba, Sydney; Kay Lehman Schlozman and Harry Brady; (1995) *Voice and Equality: Civic Voluntarism in American Politics*. Cambridge, MA, Harvard University Press.
- Virchow, Rudolf. (1985) "Report on the Typhus Epidemic in Upper Silesia. Translated", in L.J. Rather (Editor), *Rudolf Virchow: Collected Essays on Public Health and Epidemiology*. Canton, MA, Science History Publications.
- WHO (1978) *Declaration of Alma-Ata*. www.who.int/hpr/NPH/docs/declaration_almaata.pdf
- Wilkinson, Roy. (1999) "Health, Hierarchy, and Social Anxiety", *Annals Of The New York Academy Of Science*, 896, 48-63.
- Wong, Michael; Martin Shapiro; W. John Boscardin and Susan L. Ettner. (2002) "Contribution Of Major Diseases To Disparities in Mortality", *New England Journal of Medicine*, 347, 20, 1585-92.

- Infoshare Online. *Public Health Indicators*. www.infoshare.org
- MacArthur Foundation. *Supporting Creative and Effective Organizations*. www.macfound.org/research/hcd/index.htm
- Washington DC: District of Columbia. *Area Health Education Centre*. dcahec.gwumc.edu/education/index.html
- Harvard School of Public Health. *Public Health Disparities*. www.hsph.harvard.edu/thegeocodingproject/webpage/monograph/

Peter Muennig
 Mailman School of Public Health
 Columbia University
 New York, USA
 pm124 (@) columbia.edu

Websites

- World Health Organisation. www.who.int
- Gapminder World. www.gapminder.org

Corporate Governance

Eva E. Tsahuridu

Introduction

Corporate governance examines the relationships between the different primary participants of corporations, which generally include the shareholders, management and the board, in determining the relationship and performance of corporations (Monks & Minow 2001). In addition to those participants, who Monks and Minow call the 'tripod' of corporate governance, employees, customers, suppliers, creditors and others in the community are also increasingly considered as participants in the governance process.

Gompers, Ishii and Metrick (2001) describe corporate governance as 'the rules and regulations that address the problem of control, and the power that results from the separation of management and ownership in publicly traded corporations'. Gregory (2001:438) explains that the primary concern of corporate governance is "to ensure the means by which a firm's managers are held accountable to capital providers for the use of assets", while Grant (2003:925) explains that corporate governance is 'a broad theory concerned with the alignment of management and shareholder interests'. Moving beyond shareholders, The World Bank describes its activities on corporate governance as focusing "on the rights of shareholders, the equitable treatment of shareholders, the treatment of stakeholders, disclosure and transparency and the duties of board members" (World Bank 2002), emphasizing the general obligations of an organisation towards stakeholders. Similarly, the Organization for Economic Co-operation and Development (OECD) Principles of Corporate Governance (2004) refer to corporate governance as "the rules and

practices that govern the relationship between the managers and shareholders of corporations, as well as stakeholders like employees and creditors". These rules, which rely on legal, institutional and regulatory frameworks, attribute growth, financial stability, financial market integrity and economic efficiency to good corporate governance. The OECD Principles cover the issues of institutional investors, shareholder and stakeholder rights, conflicts of interest, auditor responsibilities, whistle blower protection and the responsibilities of the board.

Commonly, corporate governance addresses the interests of shareholders, and aims to guarantee their protection and promote their self-interest, by limiting the self-interested behaviour of their agents, the managers of organisations. Increasingly, however, corporate governance develops as the mechanism that enables the balancing of stakeholder interests (see e.g. Monks & Minow 2001; Siebens 2002), indicating a shift from the shareholder view of the firm to the broader stakeholder view of the firm. This shift has enabled the broadening of the focus of corporate governance, thus increasing its impact and potential effectiveness, but also increasing its complexity.

Examining corporate governance in the newly developed stakeholder emphasis, different levels of analysis can be identified. These are the societal level, the corporate level and the managerial level (Mackenzie 2004). At the societal level, corporate governance addresses how society can persuade companies to serve the public interest. At the corporate level it looks at ways of inducing the company to serve shareholders' interests and other stakeholders' interests. At the management level corporate governance attempts to find ways for management to induce behaviour from employees that serves the company's

interests. The interests and power relations between the groups at all levels of analysis, and within these groups, are not necessarily congruent. These incongruencies give rise to governance problems. The most attention given by legal and regulatory prescriptions on governance is at the corporate level where it primarily addresses the relationship between management and shareholders. Increasingly, at this level, the relationship is extended beyond the shareholders to other stakeholders. The corporate level of governance is also related to the other levels of governance, the societal and management, as both these levels inform the means and ends of corporate existence, that is they specify the purpose of the corporation (objective or goal of corporate existence) and by what means it can achieve that purpose (activities and processes).

Theories of Governance

The Organization for Economic Co-operation and Development (OECD 2004:14) in its Principles of Corporate Governance proclaims that ‘There is no single model of good corporate governance’. McKinsey (2000), however, describes a well governed company as ‘having the majority of outside directors on the board with no management ties; holding formal evaluations of directors; and being responsive to investor requests for information on governance issues. In addition, directors hold significant stockholdings in the company and a large proportion of directors’ pay is in the form of stock options’. Monks (2001) comments that in order to determine the quality of corporate governance of a company, it is important to look at the incentive structure for the CEO, in order to ascertain what the board promotes through incentives to be achieved by the CEO.

Several theoretical perspectives exist on corporate governance, these are: agency,

stewardship, stakeholder, resource dependence, management hegemony and class hegemony (see Clarke and Clegg 1998). Clarke and Clegg suggest that the two fundamental questions of corporate governance are: “Who has control?” And, “For whom is control exercised?”.

Dixon and Dogan (2002) identify four contrasting perspectives of corporate governance with incompatible contentions about corporate governance processes, arguing that these processes form an environment in which governance failures occur. Such failures lead to two possible outcomes: trench warfare between governors and those they seek to govern, which results in the victory of one over the other, or in confrontation and integration of competing governance interests and desires.

Shareholder and Stakeholder Views

Different views of corporate governance represent varying understandings to the important questions of the nature of business organisations and the main beneficiaries of the activities of business organisations. The two main views are the shareholder and stakeholder view of the firm. From the definitions of corporate governance outlined earlier and the developments in corporate governance, a shift is evident towards the stakeholder view of the firm as a more acceptable perspective at least in rhetorical terms.

The shareholder view is based on private property rights. It assumes that the corporation exists for the benefit of the shareholder and as a result, the corporations’ decisions and actions must focus on the creation of value for its shareholders. This view is founded primarily on neoclassical economics and agency theory. It assumes that behaviour motivated purely by self interest will lead to overall benefit with the assistance of the market’s invisible hand. The invisible

hand is supposed to be the market mechanism that aligns private and public interests, thus eliminating governance problems and also eliminating the need for moral motivation for behaviour. Markets are not perfect, however, and the intervention of the invisible hand is misconstrued (Stovall, Neill & Perkins 2004). The stakeholder view describes corporations as a constellation of different stakes, and suggests that corporations should act in ways that create value for all those stakes. The value created for each stake must be proportional to the contribution made by each, making this view concerned with distributive and procedural justice.

The realisation that societies and humans do not flourish as a result of self interested behaviour but as a result of concern for others is reintroduced with the stakeholder view and its increased prominence. In addition, the domination and emphasis on the primacy of the shareholders and the risk they assume in providing capital to the corporation is questioned, leading to the recognition that other stakeholders, such as employees or suppliers, may in fact assume greater risk in their relationship with the corporation. Stakeholder theory argues normatively and empirically that there is a wide web of relationships with a stake in the governance of corporations (Collier and Roberts 2001). It is also argued (see e.g. Kelly 2003) that most shareholders do not participate directly in the market, thus the notion of shareholdings is inaccurate. Kelly argues that the notion of private property rights in corporate governance is comparable to the notion of aristocracy, in that it supports the myth of the principle's right to profit. She argues that it is not boards of directors or management which govern the corporation but the share market, because any falls that are perceived as unacceptable will oust either the board, the management or both, leading to a takeover. The claim that shareholders, by holding

management accountable, act for their own benefit as well as the benefit of society is also questioned (Deakin 2005; Kennedy 2000). Deakin argues that shareholder primacy is not a legal requirement but is an outcome of the norms and practices that resulted from the hostile takeovers in the UK and US, in the 1970s and 1980s. Shareholder primacy, Deakin argues, enabled continued public evaluation via the capital market, but the information available to the market and the nature of the market itself need questioning. Management, in this context, undertake short-term share price maximisation strategies which do not necessarily benefit society and provide erratic results for shareholders.

Agreement between the shareholder and stakeholder views exists in that they both perceive the business organisation as a vehicle for the creation of value. Important differences exist, however, in the definition of that value and its beneficiaries. The shareholder view sees value as economic, profit or share price maximisation, and its beneficiary is the shareholder of the corporation. Friedman's thesis (1970) on the purpose of the business organisation represents this view. On the other hand, the stakeholder view sees value creation as a sustainable contribution to society's flourishing (Freeman 1984; Phillips, Freeman & Wicks 2003). Corporate governance and attempts to reform it are related partially to these conceptual developments, based on the realisation that business organisations affect not only their shareholders but additional stakeholders. Recently, however, the wider participation in the share market and share ownership, especially in the US, alters the clear distinction between shareholders from other stakeholders. An increasing tautology develops as most people in the US and to a lesser extent Europe, through pension and savings schemes, are shareholders as well as being members of other stakeholder groups.

The Harvard Law Review (2004) clarifies this new development and comments that many companies in the US shifted from defined benefit retirement plans to defined contribution plans. Defined contribution plans provide employees with discretion over the investment decisions of tax-deferred payments made to them during their employment. As a consequence of this shift, more than 52 million American households now own shares. They represent more than fifty percent of total American households who saw an increase in their share holdings from about \$13,000 in 1992 to over \$34,000 in 2001.

The recent corporate failures and unethical behaviour, as well as increased shareholder awareness and activism, led to increased concern about the type of value business organisations should produce and for whom. The idea that organisations exist to maximise shareholder value is challenged by differentiating between the motive and purpose of business (Duska 1997), and stakeholder theory and research (Freeman & Phillips 2002). It is argued that the purpose of business cannot be the creation of wealth but the provision of goods and services that society desires, with wealth creation being the motivation for participating in the process. The corporation in the 21st century is increasingly expected to strive for sustainable value creation, where sustainability is “the possibility that humans and other forms will flourish on the earth for ever. Flourishing means not survival, but also the realization of whatever we as humans makes life good and meaningful, including notions like justice, freedom, and dignity” (Ehrenfeld in Wheeler et al. 2003:17). Rights in general and employee rights in particular are also mentioned as an important and necessary contribution to the future of corporate governance.

The shareholder and stakeholder views of corporations affect organisational governance in terms of purpose and approach. The prevailing approach remains the shareholder view of the firm, with the emphasis on ensuring that public organisations look after the interests of their shareholders. Consequently, the ordinary understanding of corporate governance is concerned with the management of public organisations. Corporate governance addresses appropriate processes and mechanisms to ensure that the interests of the providers of capital are protected. This is fundamentally based on the agency relationship of management, where the ‘agent’ behaves in ways that benefit the shareholders, the ‘principals’.

Ongoing corporate and managerial misbehaviour, however, which results in the suffering of organisations and their stakeholders, leads to continuous attempts to reform corporate governance. Corporate misbehaviour can be described as attempts by the organisation to use illegal and/or unethical means to benefit, ultimately to the cost of its stakeholders. Managerial misbehaviour includes such action as management using corporate resources and their roles in the corporation to further their own interests to the cost of the corporation and its stakeholders.

Ultimately both views of corporations provide prescriptions to them so the questions for whom and how, can be answered. Corporate activity in general and these questions in particular are fundamentally ethical because they are about human interaction which gives rise to ethical issues and makes morality necessary so that the social group can be preserved (Emler & Hogan 1992).

Corporate Governance, Agency and Ethics
Corporate governance, in its most common conception, exists when capital is separated

from its owners, where agents manage the capital of principals. In the business context, agency theory addresses the duties of an agent to another party. At the corporate level of analysis, which represents the most popular understanding of governance, this party is the shareholder. Corporate governance is generally perceived as an economic and legal issue and not as an ethical issue, despite the fundamental ethical content of corporate activity. Attempts to resolve governance problems by legal compliance and a market approach are proving both ineffective and ineffectual, whilst they simultaneously increase the cost of business. Market failure is actually the source of corporate governance problems and the governance structure is used as a control tool, available to shareholders and government agencies to limit the discretion, and possible selfishness of directors and managers.

The corporate governance perspective that is based on agency theory, where one person (the agent) acts for another (the principal), in principle contains no ethics and therefore poses no ethical problem (Collier & Roberts 2001) or, as De George (1992) claims, it is ethically neutral, and is concerned primarily with ensuring the least costly compliance of the agent to the principal. The emphasis of corporate governance through the agency perspective focuses on self-interested individuals who need to be given incentives that promote self-interest or are controlled by legal and structural processes. This conception of governance is based on economics and the law and excludes ethics. As a consequence, it may continue to fail, because it does not take into account the defining element of human nature, morality. This conception also raises questions about the market fundamentals upon which business activity is built and their ability to promote a sustainable society or the pursuit of happiness. Corporate governance and

attempts to reform it need to address the fundamental and ontological question of what a business organisation is. The classical governance model, which insists on economic shareholder value, distorts organisations and persons. It distorts persons because it assumes that people have only self-interested motivations and fails to account for persons' moral motivations. It also distorts organisations and instead of entities that exist to assist people to pursue what is important in life, converts them to entities which takeover people's lives. Denhardt (1981:32) expresses this distortion by stating that "we originally sought to construct social institutions that would reflect our beliefs and values; now there is a danger that our values would reflect our institutions". These distortions can be explained by what Wagner-Tsukamoto (2003) calls the confusion of the modelling of reality, that is the creation of the economic man as a crash dummy so that knowledge can be created and tested, with the depiction of that reality, assuming the dummy is the man.

Even in the agency relationship however there are moral prescriptions that can address some of the governance issues discussed. In principal-agent relations De George (1992) describes three applicable principles:

- Agents are not ethically allowed to do what the principals are not ethically allowed to do.
- Agents cannot exonerate themselves for unethical actions because they are acting as agents for principals. Agents are responsible for the actions they perform, whether they are under command or on behalf of another.
- The principals are morally responsible for the actions of their agents. Agency involves the delegation of authority but not the complete delegation of (or abdication from) responsibility.

The agency relationship in this sense does not define the moral relationship, but takes place in the moral milieu (Bowie & Freeman

1992) that is it takes place in the plane of morality and as a result it is subject to moral values and moral evaluation.

Agency Problems

Agency problems in corporate governance generally describe management behaviour that does not promote shareholder value but, instead, benefits the agent, i.e. management. Agency problems are present in corporations because the relationship is assumed to exist outside the moral milieu. As a consequence, principles and agents rely primarily on imposed rules to clarify the nature and process of their relationship. The relationship between management and shareholders cannot however be fully prescribed, because management needs power and discretion to manage, and thus autonomy to do so. Agency costs exist because management does not bear the financial cost of its decisions, and also because dispersed shareholders cannot effectively monitor management (Jensen & Meckling 1976). Agency problems arise when management's decisions and behaviour benefit it rather than its principals, the shareholders of the firm. Agency problems that have been identified in the literature (see Dockery, Herbert et al 2000) include empire building, where management increases the size of the firm in order to increase its power and control through the diversification of the company's portfolio, and the problems of shirking and risk aversion.

The problems of agency result in shareholders not gaining as much growth or return of capital, as could have been the case if they were not present. The existence of the phenomenon is not surprising given the characteristics of the corporate world. This phenomenon is partially a consequence of the fallacy of Friedman's thesis, which claims that shareholders are motivated entirely by self-interest and seek to maximise profits, but managers are not (Grant 1991). Managers, the

agents of the self-interested individuals, are either totally devoid of self-interest or their self-interest is contained, allowing them to be dedicated to the self-interest of the shareholders. Agency costs and the behaviour of agents loudly contradict these assumptions.

Different Systems of Corporate Governance

The different legal and business contexts give rise to different models of corporate governance. In the literature, two clearly discernible broad systems are identified: the outsider system and the insider system (Mayer 1994; Ooghe & De Lange 2002). The outsider system which relies on regulations to ensure shareholder interests are protected, is epitomised by the Anglo-American model (including Australia and New Zealand), while the insider system which relies on shareholders' control of management is represented by Japan, Germany and other West European countries.

The existence of the two systems is a consequence of the development of corporate governance in different contexts. In the US, for example, it developed in response to economic development, liberalism and individual rights, which gave management of corporations power and influence. In addition, due to the dispersed body of shareholders in the US, the board of directors is tasked with monitoring management, thus giving the board additional power and responsibility. In Germany and Japan, however, where concentrated and cross shareholdings are allowed and are generally the norm, these large shareholders control and assist management internally (Lashgari 2004). Managers in these countries have the responsibility to act on behalf of the shareholders to provide them with appropriate returns. As a consequence, corporations in much of Europe and Japan are controlled by powerful shareholders who influence and

exercise control over management. Conversely, in the US the system relies on the law to protect shareholders from management and to provide the framework for that relationship.

The main differences between the Anglo-American and the Continental European models of corporate governance are a low concentration of shareholders in the Anglo-American context (where a large number of shareholders possess small holdings), while in continental Europe shareholder groups hold large blocks of shares (Ooghe & De Lange 2002). Shareholder identity is another notable difference. More than 50 percent of shares in the UK and the US are traded through agents of financial institutions and only 20 to 30 percent through private persons. This contrasts sharply with Continental Europe where shares are held by private companies, financial institutions and private persons. Regulations in the UK and US compel many financial institutions, which are not allowed to own shares in public companies on their own behalf, to act mainly as agents in trading shares. In continental Europe private persons and companies do not use agents to control their shareholdings, but manage them personally. Another difference between the two contexts is the liquidity of the market. In the UK and the US the number of publicly traded companies as a percentage of total companies is higher than that of Continental Europe, resulting in little personal contact between the traded company and the shareholder in the former context.

There is also a marked contrast between pension systems in the Anglo-American and Continental European models that affects shareholders and shareholdings. The UK and US provide financial resources that are managed by financial institutions. This is not the case in continental Europe where most companies are private and investment is made personally, without financial intermediaries.

A stronger personal relationship between the investor and the management of companies is thus more likely in continental Europe where, in many cases, the management is also a shareholder. The last important difference identified by Ooghe and De Langhe (2002) is that of block and mutual shareholdings. Blockholdings are higher in Europe, giving rise to a high degree of concentration of voting power. The large number of mutual holdings and the right to limited information disclosure make ownership structures in Europe less transparent than those in the Anglo American countries.

The differences between the Anglo-American and continental European systems lead to diverse degrees of power in corporations and time orientation for corporate objectives. In the UK and US, most shareholders do not have significant power due to their low concentration and, as a consequence, management is empowered with making decisions on behalf of the corporation. In many instances these decisions are motivated by personal gain rather than the benefit of the corporation and/or its shareholders. In many cases, overinvestment occurs which involves the enlargement of the corporation (empire building) to enlarge management's power base. Such corporate enlargement offers little or no benefit to the shareholders and creates conflicts of interest between them and management. In continental Europe controlling shareholders have the power to affect corporate strategy and decisions, so they are more likely to prevent phenomena such as empire building. Conflicts of interest exist, however, between controlling and minority shareholders, with little protection offered to the powerless minority shareholders. The time orientation is affected by the form and frequency of communication between management and shareholders. In the UK and US, the ability and likelihood of

shareholders to sell their shares if adverse information becomes available, provide a short-term orientation. This is another difference from the continental European model, where dominant shareholders are more likely to make decisions that improve the long-term profitability of the firm, rather than the short-term focus more prominent in the UK and the US. The UK and US models appear to be more efficient because firms have a greater availability of resources due to the increased liquidity of the market. However, the short-termism and egocentric behaviour by management create a context that seems to provide some explanation for many recent corporate scandals.

Despite the similarities between the UK and the US, there are also important differences. In the US, for example, the company president is also the CEO of the company. This is not the case in the UK where the Combined Code suggests that these roles are separated (Keenan 2004). As a result, the CEOs of UK companies appear to have less power and influence, which is also moderated by the composition of the board. Another difference is that companies in the UK may choose not to comply with the Combined Code if they can explain satisfactorily why they choose not to do so.

In addition to the countries mentioned above, there are numerous others who are in the process of developing their corporate governance systems. These countries are distinguished between the emerging and developing markets. Emerging markets are countries that are quickly entering the world business system, and includes some Eastern European countries, some Asian countries such as parts of China, Malaysia, Thailand, Indonesia, and the Philippines and some Latin American countries (Millar, et al. 2005). The developing markets include countries such as Laos, Cambodia and India. African countries are also in the process of establishing their

corporate governance structures. Generally, the alternatives discussed, concentrate on the insider or outsider systems of governance (see, e.g., Rwegasira 2000) with allowances made for adaptation to the particularities of different contexts.

Regulatory and Legal Developments

The revised OECD Principles of Corporate Governance, ratified by the OECD countries in 2004, provide an international benchmark for corporate governance and the basis for corporate governance reform (OECD, 2004). The OECD Principles of Corporate Governance (OECD 2004) were issued in 1999 and revised in 2003. These principles cover six main areas: ensuring the basis for an effective corporate governance framework; protecting shareholder rights; ensuring the equitable treatment of stakeholders; recognising shareholders' rights; timely and accurate disclosure and transparency; ensuring the board fulfils its responsibilities.

United Kingdom

The corporate governance movement, which started in the mid-1980s, is aimed at reinstating ownership in the governance role in organisations by making boards and executives accountable to their owners (Carlsson 2001). In the UK, the Combined Code on Corporate Governance resulted from significant corporate scandals, which were followed by the Cadbury, Greenbury and the Hampel and Turnbull reports (Keenan 2004). The 1992 report of the UK Cadbury Committee is recognised as a landmark in the development of corporate governance (Carlsson 2001; Vinten 1998). The Cadbury Report provided a code of best practice addressing the board of directors, non-executive directors, executive directors, reporting and controls (Cadbury Report 1992).

The London Stock Exchange requires its listed companies to comply with the Combined Code on Corporate Governance and they are expected to issue a statement of compliance with the Code. The Greenbury report (1995) addressed the issue of executive compensation and provided a code of best practice. The 1998 Hampel Committee was critical of Cadbury for its preoccupation with accountability in public debate: "... the emphasis on accountability has tended to obscure a board's first responsibility – to enhance the prosperity of the business over time." (Hampel Committee 1998:7, cited in Spira 2001). Spira sees a schism in subsequent analyses between the view that good corporate governance depends on both enterprise (growth) and accountability, and that accountability concerns may inhibit enterprise. The current reviewed Combined Code is an outcome of further reviews of the Combined Code, issued originally by the Turnbull Report. These reviews were undertaken by the Hampel Committee, with Derek Higgs on the role of non-executive directors, and a review of audit committees by Sir Robert Smith (see Financial Services Authority 2003 for the Combined Code).

The UK's Combined Code, which has been effective since the 3rd of November 2003, addresses internal controls, including financial, operational, compliance and risk management and suggests that private, public or government organisations need to follow the Code, in order to be good corporate citizens. It is a best practice code and not a code that sets minimum requirements. The Combined Code distinguishes between principles and provisions. It allows companies to choose how they will implement the principles that the code prescribes. It requires companies to explain in their annual report how they have implemented the Code's principles, and to

further explain failure to comply with any of the provisions.

Legal reforms currently under way in the UK, following a three year review, aim to increase investor confidence and market efficiency. The Company Law Review aims to deliver many improvements, to enable better shareholder engagement, modernise and de-regulate the law, clarify directors' duties and responsibilities, facilitate better communication with shareholders and provide mechanisms for improved decision making (Department of Trade and Industry, 2004).

United States

Following the last wave of corporate and managerial misbehaviour, which occurred at the turn of the millennium, with the US as its epicentre, both legislation and regulations have changed in an attempt to prevent similar occurrences in the future. It is worth noting, however, that, in all cases, legislative and regulatory coverage existed but was not complied with. The new efforts are based primarily on legal/compliance attempts to force organisations and their management to behave in ways that protect the interests of shareholders. The argument exists that these attempts increase the inefficiency of the market by increasing the cost of doing business without guaranteeing protection.

Grant (2003) explains that in the US, legislation to protect shareholders follows share market crashes. The Securities Acts of 1933 and 1934 followed the share market crash of 1929. The Sarbanes Oxley Act followed the share market crash of 2001, which saw the corporate scandals of Enron, Worldcom and several other companies. These scandals resulted from misleading accounts and information to the market and contributed to the crash and loss of confidence of the market. The Sarbanes Oxley Act (SOA) aims to protect

shareholders by improving the accuracy and reliability of corporate information. The SOA affects governance reform not only in the US but globally, and covers areas which were in the domain of management prerogative, such as the existence of a code of ethics.

In the US, the board of directors plays an important role in the corporation, where it oversees management, sets corporate strategy and executive compensation. In practice however, the board is not independent and effective due to conflicts of interest and psychological pressures. In a recent example, a survey by Felton (2004) found that despite progress made on corporate governance reform, directors and institutional directors require further reform in separating the roles of the board, improving its accountability and addressing the issue of executive compensation.

The Sarbanes-Oxley Act, Section 404, requires companies to file a management assertion and auditor attestation on the effectiveness of internal controls over financial reporting. This requirement relates to appropriate internal control and compels management (chief executive officers and chief financial officers) to take responsibility for its establishment and maintenance and to report on the effectiveness of those controls. To comply with this section, the company annual report needs to include a statement on management's responsibility for maintaining internal controls and financial reporting procedures. The annual report must also contain an assessment of those controls and procedures, which must further be attested to and reported on by the company's outside auditors (see Corporate Directors Forum, 2003 for Sarbanes Oxley requirements and details). Sarbanes-Oxley, Section 409, requires real time disclosures of material changes in a company's financial position.

It is estimated that the SOA is costing US companies, with increased expenditure of

compliance of over 1000 percent between 2003 and 2004. It is estimated that the cost to large companies in America will exceed, on average, US\$35m in 2005 (Henry et al 2005). There is no exact cost for all companies that need to comply with SOA, but there is a consensus that the costs of compliance are rising and increasing the financial burden on companies. The costs of the new legislation and regulations include (D'Aguila, 2004): increased accounting and audit fees in internal and external hours of work and additional audit fees, training, technology, and development of policy; increase in cost of board of directors and auditors in terms of time, liability insurance and external consultancy fees. Indirect costs of the new regulations include increases of the costs of going public by 100%, the effect of the regulations and their compliance on decision making and productivity and the cost of the independent director. Small and medium size companies are more likely to be affected by the new regulations and the cost of compliance, because they are least likely to be able to comply with the changes. Another consequence of the new regulations has been an increase in the privatisation of public enterprises. Grant Thornton LLP reports that the time and cost of complying with the SOA is the reason why many public companies are privatising (cited in Corporate Governance 2004). The impression also exists that the SOA and related regulations are affecting not only public companies, but the whole business community. This is because they establish best practice, which private companies are forced to follow, in order to satisfy financial and government institutions.

In addition to the changes that resulted from the introduction of the Sarbanes-Oxley Act in 2002, the New York Stock Exchange (NYSE) and NASDAQ Listing Standards have also been changed. The new standards require that the majority of directors in a

listed company be independent, which is considered an important attribute of good governance. This independence is, however, described only very broadly in the regulations and, as a result, the effectiveness of these measures is dubious. Empirical research on the relationship between structural independence of the board of directors (such as the separation of chair and CEO, having a majority of external independent directors and the size of the board) and the corporation's financial performance is mixed, at best (Leblanc 2004). Leblanc argues that, despite the lack of solid empirical evidence to support a positive relationship between board governance and performance, directors overwhelmingly support the view that such a relationship does, in fact, exist. The methods of measurement, however, are inadequate to capture the characteristics of effective boards. Leblanc argues that it is necessary to understand and research the process as well as the structure of governance, as the actual process is the important element that can provide the link between board governance and performance.

The new regulations address not only financial conflicts, as was the case with previous attempts, but, more importantly, they also address psychological conflicts. The new regulations impose new structural and procedural requirements, such as meetings of the board from which the corporation management is excluded, thus enabling directors to be more vocal (see Harvard Law Review 2004). The Harvard Law Review explains that the new regulations provide 'legal cover' to directors by requiring them to behave in ways that make their jobs easier and more effective, without alienating management in the process. The realisation that directors are susceptible to the psychological limitations of a group is likely to make current and future reform attempts more successful.

The proliferation of corporate governance regulations, and codes as well as shareholder and market interest, have led to the development of attempts to assess or rate companies in terms of their corporate governance. In the US, Standard & Poor's, who provide investment research and credit ratings, have started granting Corporate Governance Scores (CGS) to companies on a scale from 1 (lowest) to 10 (highest). These scores, it is argued, reflect actual governance practices and not regulatory or legal compliance and, as a result, are globally relevant and comparable.

Resolving Governance Problems

The failure of corporate governance has been attributed to a number of causes, among them lack of knowledge of governance, lack of governance instruments, lack of governance effectiveness and motivation problems, as expressed by lack of compliance by the governed (Dixon & Dogan 2002). There is no consensus on the resolution to the problems of governance as there is no consensus on what is good governance. The two generally accepted views see governance problems resolved either through the market mechanism or by legislation (Harvard Law Review 2004). The market view perceives regulation unnecessary and inefficient. It sees the market as sufficiently able to efficiently and effectively resolve governance issues, and result in appropriate governance structures. The legislative view sees regulation necessary, because it finds it improbable that corporations will voluntarily adopt governance structures that are acceptable by shareholders. It is thus necessary to introduce regulations to insure confidence and trust in the market. A third view, which is the least popular and the least discussed, is the ethics or values based approach (Collier & Roberts 2001; MacKenzie 2004). It sees the remoralisation

of business, or the end of the separation thesis between ethics and business, as a necessary condition for the resolution of governance problems. It argues that it is necessary to question the purpose of corporations and the means of achieving that purpose. It does not rely on compliance to regulations or market forces, but on ethical awareness and behaviour. As such, it is more likely to be effective and least costly to implement. Such an approach requires a more holistic examination of not only markets and corporations but also governments and politics.

Selected References

- Bowie, N. E., and Freeman, E. R. (1992) "Ethics And Agency Theory", in N.E. Bowie and E. R. Freeman (Editors), *Ethics And Agency Theory: An Introduction*. Third Edition. New York: Oxford University Press, pp. 3-24.
- Cadbury Report. (1992) *Report of the Committee on the Financial Aspects of Corporate Governance*. London: Gee Publishing.
- Carlsson, R. H. (2001) *Ownership and Value Creation: Strategic Corporate Governance in the New Economy*. Chichester: John Wiley.
- Clarke, T., and S.R. Clegg. (1998) *Changing Paradigms: the Transformation of Management Knowledge for the 21st Century*. London: Harper Collins Business.
- Collier, J., and J. Roberts. (2001) "An Ethic for Corporate Governance?", *Business Ethics Quarterly*, 11(1), 67-71.
- Corporate Directors Forum. (2003) *Status of Sarbanes Oxley Requirements and Related Sec, Nasdaq and Nyse Proposals*. Corporate Directors Forum. www.directorsforum.com/cgnews_soxstat.us.Html
- Corporate Governance. (2004) "Corporate Governance Update", *Corporate Governance*, 12(3), 408-414.
- D'Aguila, J.M. (2004) "Tallying The Cost Of The Sarbanes-Oxley Act", *CPA Journal*, Nov., 6-9.
- Deakin, S. (2005) "The Coming Transformation Of Shareholder Value", *Corporate Governance: An International Review*, 13(1), 11-18.
- De George, R.T. (1992) "Agency Theory And The Ethics Of Agency", in N.E. Bowie and E.R. Freeman (Editors), *Ethics And Agency Theory: An Introduction*. Third Edition. New York: Oxford University Press, pp. 59-74.
- Denhardt, R.B. (1981) *In the Shadow of the Organisation*. Lawrence, KS: Regents Press.
- Department of Trade and Industry. (2004) *Modernising Company Law*. www.dti.gov.uk/cld/review.htm
- Dixon, J., and R. Dogan. (2002) "Hierarchies, Networks and Markets: Responses to Societal Governance Failure", *Administrative Theory and Praxis*, 24(1), 175ff.
- Dockery, E.; W.E. Herbert and K. Taylor. (2000) "Corporate Governance, Managerial Strategies and Shareholder Wealth Maximisation: A Study of Large European Companies", *Managerial Finance*, 26(9), 21-35.
- Duska, R. F. (1997) "The Why's of Business Revisited", *Journal Of Business Ethics*, 16, 1401-1409.
- Emler, N. and R. Hogan. (1992) "Individualizing Conscience: New Thoughts on Old Issues", in W.M. Kurtines, M. Azmitia and J.L. Gewirtz (Editors), *The Role Of Values In Psychology And Human Development*. New York: John Wiley, pp. 200-238.
- Felton, R.F. (2004) "What Directors and Investors Want from Governance

- Reform”, *Mckinsey Quarterly* 2. www.mckinseyquarterly.com
- Financial Services Authority. (2003) *The Combined Code On Corporate Governance*. Financial Reporting Council. www.Fsa.Gov.Uk/Pubs/UKla/Lr_Comcode2003.Pdf
- Freeman, R.E. (1984) *Strategic Management: A Stakeholder Approach*. Boston: Pitman.
- Freeman, R.E. and R.A. Phillips. (2002) “Stakeholder Theory: A Libertarian Defence”, *Business Ethics Quarterly*, 12(3), 331-349.
- Friedman, M. (1970) “The Social Responsibility of Business is to Increase its Profits” in W.M. Hoffman and J.M. Moore (Editors), *Business Ethics: Readings and Cases in Corporate Morality*. New York: McGraw-Hill, pp. 126-131.
- Gompers, P. A., Ishii, J. L., and Metrick, A. (2001) *Corporate Governance and Equity Prices*. NBER Working Paper No. W8449. New York: NBER.
- Grant, C. (1991) “Friedman Fallacies”, *Journal Of Business Ethics*, 10, 907-914.
- Grant, G.H. (2003) “The Evolution of Corporate Governance and its Impact on Modern Corporate America”, *Management Decision*, 41(9), 923-934.
- Greenbury, R (1995) *Directors Remuneration: Report Of A Study Group Chaired By Sir Richard Greenbury*, London: Gee.
- Gregory, H. (2001) “Overview of Corporate Governance Guidelines and Codes of Best Practice in Developing and Emerging Markets”, in R. Monks and N. Minow (Editors), *Corporate Governance*. Oxford: Blackwell Publishers, pp. 438-446.
- Harvard Law Review (2004) “Developments in the Law: Corporations and Society”, *Harvard Law Review*, 117(7), 2169-2295.
- Henry, D.; A. Borras; L. Lavelle; D. Brady; M. Arndt and J. Weber. (2005) “Death, Taxes and Sarbanes-Oxley?”, *Business Week*, 6 January, 28-31.
- Jensen, M.C., and W.H. Meckling. (1976) “Theory of the Firm: Managerial Behaviour, Agency Costs and Ownership Structure”, *Journal Of Financial Economics*, 305-360.
- Keenan, J. (2004) “Corporate Governance in UK/US Boardrooms”, *Corporate Governance*, 12(2), 172-176.
- Kelly, M. (2003) *The Divine Right Of Capital*. San Francisco: Berrett-Koehler.
- Kennedy, A. (2000) *The End of Shareholder Value: Corporations At the Crossroads*. Cambridge, MA: Perseus.
- Lashgari, M. (2004) “Corporate Governance: Theory and Practice”, *Journal Of American Academy Of Business*, Cambridge, 5(1/2), 46-51.
- Leblanc, R.W. (2004) “What's Wrong With Corporate Governance: A Note”, *Corporate Governance: An International Review*, 12(4), 436-441.
- Mackenzie, C. (2004) “Moral Sanctions: Ethical Norms as a Solution to Corporate Governance Problems”, *Journal Of Corporate Citizenship* 15, 49-61.
- Mckinsey & Co. (2000) *Investor Opinion Survey*. www.oecd.org/dataoecd/56/7/1922101.pdf.
- Mayer, C. (1994) “Stock-Markets, Financial Institutions, and Corporate Governance”, in N. Dimsdale and M. Prevezer (Editors), *Capital Markets And Corporate Governance*. Oxford: Clarendon Press, pp. 179-194.
- Millar, C.C.; T.I. Eldomiaty; C.J. Choi and B. Hilton. (2005) “Corporate Governance and Institutional Transparency in Emerging Markets”, *Journal of Business Ethics*, 59, 163-174.
- Monks, R.A.G. (2001) “Redesigning Corporate Governance Structures And Systems For The Twenty First Century”,

- Corporate Governance: An International Review*, 9(3), 142ff.
- Monks, R.A.G. and N. Minow. (2001) *Corporate Governance*. Malden, MA: Blackwell.
- OECD. (2004) *The OECD Principles of Corporate Governance*.
www.oecd.org/dataoecd/32/18/31557724.Pdf
- Ooghe, H., and T. De Langhe. (2002) "The Anglo-American Versus the Continental European Corporate Governance Model: Empirical Evidence of Board Composition in Belgium", *European Business Review*, 14(6), 437-449.
- Phillips, R.; R.E. Freeman and A.C. Wicks. (2003) "What Stakeholder Theory is Not", *Business Ethics Quarterly*, 13(4), 479-502.
- Rwegasira, K. (2000) "Corporate Governance in Emerging Capital Markets: Whither Africa?", *Corporate Governance*, 8(3), 258-267.
- Siebens, H. (2002) "Concepts and Working Instruments for Corporate Governance", *Journal of Business Ethics*, 39(1), 109-116.
- Spira, L.F. (2001) "Enterprise And Accountability: Striking A Balance", *Management Decision*, 39(9), 739-748.
- Stovall, O.S.; J.D. Neil and D. Perkins. (2004) "Corporate Governance, Internal Decision Making, and the Invisible Hand", *Journals of Business Ethics*, 51(2), 221-227.
- Vinten, G. (1998) "Corporate Governance: An International State of the Art", *Managerial Auditing Journal*, 13(7), 419-431.
- Wagner-Tsukamoto, S. (2003) *Human Nature And Organization Theory: On The Economic Approach To Institutional Organization*. Cheltenham, UK: Edward Elgar.
- Wheeler, D.; B. Colbert and R.E. Freeman. (2003) "Focusing on Value: Reconciling Corporate Social Responsibility, Sustainability and a Stakeholder Approach in a Network World", *Journal Of General Management*, 28(3), 1-28.
- World Bank. (2002) *About Corporate Governance*. Washington: World Bank.
www.worldbank.org/html/fpd/privatesector/cg/index.htm.

Selected Websites

Governance website.

www.governance.co.uk/index.htm.

Corporate Governance website.

www.corpgov.net/.

Global Corporate Governance Forum.

www.gcgf.org/.

Eva E. Tsahuridu
School of Management
RMIT University
Melbourne, Australia
eva.tsahuridu@rmit.edu.au

Discrimination

Shane Ostenfled

Introduction

Discrimination refers to the application of subjective criteria that results in a person being disadvantaged on the basis of characteristics ascribed to a group. Most of the debate centres on discrimination in employment. Discrimination patterns and employment disadvantage may be evidenced in employment segregation, marginal employment status, and pay inequity. Individuals may experience disadvantage arising from either direct or indirect discrimination before or after entering employment. Educational discrimination is related to employment discrimination, as access to high pay jobs requires educational qualifications.

Both neo-classical and institutional theories of disadvantage give prominence in explanation to discrimination. Neo-classical theories adopt 'human capital' theory to understand pre-entry discrimination. Post-entry discrimination, in areas such as access to and promotion to jobs, is ascribed to the 'taste' theory of discrimination. Institutional theorists have developed 'statistical' explanations of discrimination. Radical theorists suggest that the ruling classes divide the remainder of the population as a control strategy.

The common law gives no protection against disadvantage and thus discrimination and equality of opportunity have been civil rights issues. As a result governments have legislated against unfair discrimination in areas such as education and employment. Unfair discrimination is generally restricted in application to unfair discrimination on the basis of characteristics listed in government legislation. These may include age, weight, impairment (sometimes denoted by 'medical

record'), medical condition, race, sex, sexuality (sometimes denoted by 'homosexuality', 'sexual preference' or 'lawful sexual activity') or gender identification, marital or parental status, pregnancy, national origin, ethnicity, religion, political belief, membership of a trade union or a political party, and criminal record.

The following sections will refer to the forms and evidence of discrimination, theories of disadvantage, and legal remedies to discrimination in employment. Some general historical background to public policy and the matter of discrimination is first provided, along with a radical view of the background to legislation in terms of the role of the state in identifying 'minority' groups.

Historical Background and Identification of 'Minority' Groups

The 'rediscovery of poverty' (Connell and Irving 1992) in the 1960s generated political debate about equity in the 1970s and 1980s. Earlier debates on class privilege now turned to discussions focused on the most deprived groups. These groups became known as 'minorities', though this is a misnomer as women have been identified as disadvantaged as a group. This radical political economy view sees a transformation in the debates to the final decades of the twentieth century over social inequality from being about 'class' then turning to 'minorities'. This transformation is reflected in the change of discourse in the management and organisational literature. Here the 1970s and 1980s discussions of anti-discrimination and equal employment opportunity turn, in the 1990s, to discussions of 'diversity management'.

Legislative attempts to redress discrimination pre-date these debates. The fifth, thirteenth and fourteenth amendments to the *Constitution of the United States of America* give protection to all before the law,

against slavery, and against state legislation, respectively. These amendments took place between 1791 and 1868. The Civil Rights Act 1866 gave all citizens the right to make and enforce contracts. It has only been recently, however, that this Civil Rights Act provision, now known as *Title 42, Section 1981 of the U.S. Code* has been utilised in attacking discrimination. Indeed, notwithstanding such earlier enactments and constitutional amendments, it was not until the 1960s that the United States led the way in taking dramatic action on discrimination in employment.

As a precursor to this in 1948 the General Assembly of the United Nations passed the *Universal Declaration of Human Rights*. Article 23 states that “everyone has the right to work, to free choice of employment, to just and favourable conditions of work and to protection against unemployment, and...everyone, without any discrimination, has the right to equal pay for equal work...” This was followed by the International Labour Organisation’s *Discrimination (Employment and Occupation) Convention (No 111)* that was adopted in 1958, coming into force in 1960. This is one of the fundamental conventions of the ILO and gives substance to the 1944 ILO *Declaration of Philadelphia*.

Evidence of Discrimination Patterns and Employment Disadvantage

Employment segregation, marginal employment status and pay inequity provide much of the evidence of discrimination patterns and employment disadvantage.

Employment segregation can be either horizontal (employment in a narrow range of jobs) or vertical (employment in the jobs within an industry or occupation that offer the least status, pay, benefits and opportunities). Australian, for example, has one of the most segregated labour markets, with 89.7% of

trade workers being male and 78.7% of clerical workers being female (ABS 6203.0 May 1996). This reflects both horizontal and vertical segregation, with women working in a narrow range of jobs that are the lowest paying. Industry as well as occupational segregation is evident in Australia. More than 85% of the workforce in mining, utilities and construction is male, whereas women make up 76.4% of workers in community services (ABS 6203.0 May 1996). Similar statistics may be found for elsewhere. In the United States, Mitra reports research that has found up to 93.4% of jobs being segregated by sex in California. In the computer industry women are concentrated in low-level jobs that involve routine tasks and that paid low wages. White males typically hold jobs that involve high pay and status as well as opportunities for upward mobility. In Zambia, the difference in the proportion of women and men working in the formal sector is 31 percentage points, with most of this difference being a result of discrimination. The discrimination stems from education, and lowly educated females are affected (Skyt 1998).

Marginal employment status is evidenced in higher levels of unemployment, part-time and contract work, and outwork. Disadvantage is experienced through marginal employment status in access to superannuation and training and development. It is those groups that experience discrimination as a result of stigma that are most at risk of marginal employment status. U.S. studies have show, for example, that people with epilepsy have a 25% rate of unemployment, and 64% of these attribute their unemployment to epilepsy. Among people whose seizures are poorly controlled, the unemployment rate approaches 50% (Morrell 2002). Similarly, a ‘religion penalty’ means that Catholic men are over-represented amongst the jobless of

Northern Ireland, and under-represented as employees in professional/managerial occupations (Borooah 1999). In Britain, and using information pooled over economic cycles in the 1970s and 1980s, research has shown that lack of employment opportunities is a problem for ethnic minorities. The ethnic wage gap increased from 2.6 percentage points in the 1970s to 10.9 percent in the 1980s (Blackaby et al 1998).

Pay inequity results in disadvantaged groups earning less than the rest of the workforce, on average. Australia has taken great strides to reduce the gender wage gap such that by the mid-1990s women earned on average 83% of male full-time ordinary time earnings and 78% of full-time total earnings. Overall women earn 65% of male weekly total earnings when the large proportion of women in part-time employment is taken into account (Gardner and Palmer 1997).

Theory and Concepts

Orthodox economic theory suggests that the labour market is impersonal and allocates workers across different tiers of jobs based on their human capital attributes. But sociological studies emphasise that employers do not always hire and promote in accordance with the neoclassical models of profit maximisation, efficiency and rationality. Employer perceptions, beliefs, customs, and statistical discrimination play important roles in stratifying men and women across different jobs, titles, and tasks (Mitra 2003).

Orthodox neo-classical Human Capital theory predicts a 'trickling down' effect on earnings equalisation through the expansion of education. Through education expansion and employment legislation, it is understood that the state can reduce earnings differentials. However research has shown that education expansion for females has reduced gender earnings differentials only to a limit (eg Chung 1996). A systematic

undervaluation, or, 'animus discrimination', exists, that is motivated by ignorance, viciousness, or irrationality, and dehumanises people by categorisation. Developed by the neo-classicist Becker (1957) as the 'taste' theory of discrimination, this concept has been most developed within institutional theories. Institutional theories term this 'statistical discrimination'. Statistical discrimination consists of attributing to all people of a particular age the characteristics of the average person of that age. It is the failure or refusal to distinguish a particular member of a group from the average member. It is normally motivated by the costs of information (Posner 1999).

Criticisms of the taste theory include that it does not explain why there is a taste for discrimination, that discrimination can often appear to be rational economic behaviour, and that its long-term persistence means that the market is imperfect. This has resulted in a second neo-classical theory of 'collusive' discrimination. A small part of discriminatory behaviour may be explained by collusion - where control over entry to a profession or occupation may be restricted to a small number of unions or professional associations. Monopsonistic discrimination is a third theory that has been developed by neo-classicists. This theory suggests that some groups in the labour market, primarily women, have fewer employment alternatives. Cultural and family-level factors affect not only whether women are in the labour market but also their success in finding a job. Cultural prescriptions on female mobility have been found to be a significant constraint in women's job searches (Miles 2003).

The development of the theory of statistical discrimination by institutional theorists has focused on gathering evidence of stereotypes and the use of group-based statistical rules in discriminating against particular demographic groups. Turnover is

one such area of employer perceptions. Whilst both neo-classical and institutional theorists accept the theory of statistical discrimination neo-classicists usually believe that the spread of discrimination is limited by labour market competition. Institutional theory emphasises forces that deepen and widen discrimination. This has resulted in the development of the dual labour market hypothesis. This holds that as a result of technological advances and the need for firms to capture stable and skilled workers a primary labour market has developed that excludes members of disadvantaged groups because of their perceived instability. Radical critiques of the dual labour market hypothesis suggest that the development of dual markets is a result of employer attempts to achieve greater control over the labour process (Whitfield 1987).

Policy Responses

There are three types of policy responses to disadvantage: pay equity policies; equal opportunity policies; and education and training policies. The type of policy pursued depends partly on the reasons for disadvantage, and is also the outcome of negotiations between pressure groups and governments (Whitfield 1987).

In the United States many states passed laws prohibiting various forms of discrimination prior to the 1964 Civil Rights Act. Colorado, for example, was the first to out-law age discrimination when it did so in 1903 (Adams 2003). Activism by Jewish organisations, the National Association for the Advancement of Coloured Peoples, and trade unions (particularly the Congress of Industrial Organisations) was the key factor in securing the adoption of fair employment legislation in the U.S. states in this period (Collins 2003).

The prototype for the state anti-discrimination efforts came from the federal

level during World War II. A. Philip Randolph, leader of the Brotherhood of Sleeping Car Porters, formed the March on Washington Movement. He threatened to lead 100,000 blacks in a march in 1941 to protest segregation and discrimination. President Roosevelt agreed to issue an executive order to ban discrimination in the employment of workers in defence industries or government because of race, creed, colour, or national origin. These orders were not given legislative effect however.

Until the Civil Rights Act of 1964, congressional bills prohibiting discrimination in employment were routinely detained in committee. On the occasions when the bills were debated they expired through filibusters by southern senators. In these circumstances only state legislation gave protection, and under the state legislation it became unlawful for employers, unions, or employment agencies to discriminate on the basis of race, religion, or national origin in decisions concerning employment, discharge, referral, compensation, or other conditions and privileges of employment (Collins 2003).

Title VII of the 1964 Civil Rights Act added colour and sex as proscribed grounds for discrimination. It established an Equal Employment Opportunity Commission to investigate employment discrimination complaints and sue on behalf of complainants. The Civil Rights Act is augmented by executive orders issued in the Johnson administration that require employers who do business with the U.S. government to take affirmative action to ensure equal employment opportunity. The Equal Pay Act of 1963 also made it illegal to discriminate in pay on the basis of sex where jobs involve equal work (Dessler 1997).

U.S. Federal legislation on age discrimination came with the Age Discrimination in Employment Act (ADEA) of 1967 that prohibited discrimination based

on age for people aged between 40 and 65. An amendment in 1978 increased the minimum allowable mandatory retirement age to 70 and in 1986 mandatory retirement was banned at all ages. The ADEA has been used as the basis for many claims of age discrimination. There were a total of 167,959 charges resolved under this legislation between fiscal year 1992 and fiscal year 2001. In the same period there were 320,485 race discrimination charges and 274,474 sex discrimination charges resolved (Adams 2003).

Further legislation related to outlawing discrimination followed in the U.S. through the 1970s. This included the Vocational Rehabilitation Act of 1973, the Vietnam Era Veterans' Readjustment Assistance Act of 1974 and the Pregnancy Discrimination Act of 1978. Federal agency guidelines were also developed for a range of areas, including uniform employee selection procedures (Dessler 1997).

A range of court decisions helped define rights under anti-discrimination legislation and, in doing so, the U.S. courts championed the protection of women and minority groups until cases such as *Wards Cove Packing Company v. Atonio* and *Patterson v. McLean Credit Union* had the effect of limiting such protection. The U.S. Congress then passed a new Civil Rights Act in 1991. This placed the burden of proof back on employers and permitted compensatory and punitive damages. Also in the 1990s came the Americans with Disabilities Act, signed into law by President Bush in 1990.

Britain and most European and English-speaking countries followed the path laid out by the U.S. legislature. In Britain an Equal Pay Act was passed in 1970, to be followed by Sex Discrimination Acts and a Race Relations Act. The Fair Employment (Northern Ireland) Act of 1989 made it unlawful to discriminate on the basis of

religious or political belief or affiliation (Noon and Blyton 1997).

The Australian Labor Party (ALP) government in Australia between 1983 and 1996 pursued all three areas of policy response to discrimination. It introduced legislation concerning discrimination in employment and, later in its term, discussed conditions which assisted workers to combine their responsibilities in the paid workforce and within the home. While outcomes relied on individual action (for example in the case of anti-discrimination cases) or activity at the organisation level (affirmative action programs and 'work and family' policies), there was a legislative base to part of this agenda (Ostenfeld and Strachan 1999).

In Australia the *Racial Discrimination Act* 1975 was passed during the Whitlam Labor Government and this government intended to introduce a similar act making discrimination on the grounds of sex unlawful. The Liberal/National Party Coalition Government (1975 to 1983) did not pursue this and it was left to the next Labor Government to introduce both sex discrimination and affirmative action legislation. However, the federal Labor Government was not alone in pursuing this course - several state Labor Governments had introduced this legislation and these moves can be seen as part of a wider international push for legislation and policies which recognised the disadvantaged position of women in the labour force. Groups within the Australian community such as the Women's Electoral Lobby pushed for specific legislation that would recognise women's disadvantaged position in the labour market and community. The ALP promised it would introduce such laws and Senator Ryan, in her speech introducing the Sex Discrimination Bill, said that it represented an 'initial step towards the fulfilment of the Government's major election commitment to women'. The *Sex Discrimination Act* 1984

aroused a degree of public furore and condemnation. Attacks on the Bill came from some of the churches and conservative women's groups such as 'Women Who Want to be Women'. The *Affirmative Action (Equal Opportunity for Women) Act* was introduced in August 1986 following similar provisions in the *Public Service Reform Act 1984* and legislation relating to some state public service workers (Ostenfeld and Strachan 1999).

A significant aspect of the British and Australian legislation is that it outlaws both indirect and direct discrimination. Direct discrimination occurs when a person with a particular attribute is treated less favourably, or it is proposed to treat that person less favourably, than a person without the attribute, in comparable circumstances. Indirect discrimination occurs where the characteristics of the dominant group are posited as the 'norm'. A celebrated Australian case held that the requirement for a particular length of leg to be a bus driver, for example, constituted indirect discrimination (Gardner and Palmer). The Australian legislation also treats sexual harassment as a separate category of discrimination.

In Eastern Europe legislation in most countries has long included clauses about equality of men and women. However anti-discrimination labour market policies were not introduced until recently. Joining the European Union involves a process of harmonisation of legislation, and so countries including the Czech Republic and Slovakia are now in the process of enacting policies to ensure comparable worth, equal pay, and equal employment opportunity (Jurajda 2003).

The approach to anti-discrimination and equal employment opportunity in Europe and elsewhere has not been without criticism. The critiques of the British Equal Employment Opportunities Commission are that does not

completely investigate all charges, and does not apply the law to the fullest benefit of discrimination victims. Wilhelm (2001) for example, finds that, among women, benefits from EEOC enforcement favour white women and black women with relatively high levels of education. To radical critics of liberal anti-discrimination measures it is structural disadvantage that needs to be addressed, and this needs to be undertaken through positive discrimination, or affirmative action.

Selected References

- Adams, Scott J. (2003) "Age Discrimination Legislation and the Employment of Older Workers", *Labour Economics*, 223-249.
- Blackaby, D.H.; D.G. Leslie; P.D. Murphy and N.C. O'Leary, (1998) "The Ethnic Wage Gap and Employment Differentials in the 1990s: Evidence for Britain", *Economics Letters*, 58, 1, January, 97-103.
- Borooah, Vani K. (1999) "Is There a Penalty to Being a Catholic in Northern Ireland: An Econometric Analysis of the Relationship Between Religious Belief and Occupational Success", *European Journal Of Political Economy*, 15, 2, June, 163-192.
- Chung, Yue-Ping. (1996) "Gender Earnings Differentials in Hong Kong: The Effect of the State, Education, And Employment", *Economics of Education Review*, 15, 3, June, 231-243.
- Collins, William J. (2003) "The Political Economy of State-Level Fair Employment Laws, 1940-1964", *Explorations In Economic History*, 40, 1, January, 24-51.
- Connell, R. W. and T. H. Irving. (1992) *Class Structure In Australian History*. Second Edition. Melbourne: Longman Cheshire.
- Dessler, Gary. (1997) *Human Resource Management. International Edition*. Prentice Hall, Upper Saddle River.

Gardner, Margaret and Gill Palmer. (1997) *Employment Relations: Industrial Relations and Human Resource Management in Australia*. Second Edition. Melbourne: Macmillan.

Jurajda, Stepán. (2003) “Gender Wage Gap and Segregation in Enterprises and the Public Sector in Late Transition Countries”, *Journal Of Comparative Economics*, 31, 2, June, 199-222.

Miles, Rebecca. (2002) “Employment and Unemployment in Jordan: The Importance of the Gender System”, *World Development*, 30, 3, March, 413-427.

Mitra, Aparna. (2003) “Establishment Size, Employment, and the Gender Wage Gap”, *Journal of Socio-Economics*, 32, 3, 317-330.

Morrell, Martha J. (2002) “Stigma and Epilepsy”, *Epilepsy & Behavior*, 3, 6, Supplement 2, December, 21-25.

Noon, Mike and Paul Blyton. (1997) *The Realities of Work*. London: Macmillan.

Posner, Richard A. (1999) “Employment Discrimination: Age Discrimination and Sexual Harassment”, *International Review of Law and Economics*, 19, 4, December, 421-446.

Skyt, Helena (1998) “Nielsen Discrimination and Detailed Decomposition in a Logit Model”, *Economics Letters*, 61, 1, October, 115-120.

Whitfield, Keith. (1987) *The Australian Labour Market*. Sydney: Harper & Row.

Wilhelm, Sarah. (2001) “The Impact of EEOC Enforcement on the Wages of Black and White Women: Does Class Matter?”, *Review Of Radical Political Economics*, 33, 3, Summer, 295-304.

US House of Representatives. *Constitution*.
www.house.gov/Constitution/Constitution.html

Shane Ostenfeld
Faculty of Business and Law
University of Newcastle
New South Wales, Australia

Websites

Universal Declaration of Human Rights.
www.un.org/Overview/rights.html

Discrimination Convention 1958.
www.ilo.org/ilolex/cgi-lex/convde.pl?C111

Economic Growth and Environment

Oscar Alfranca

Introduction

Although the main insights of the relationships between economic growth and the environment can be found in the research of Thomas Malthus, John Stuart Mill and Stanley Jevons, it is in the debates during the 1970s on “the limits to growth” (Meadows et al 1972), that the relationship between economic growth and the environment is discussed “holistically” (Edward-Jones, Davies et al 2000). In the wake of the oil market shocks in the 1970s, some central works such as Dasgupta and Heal (1974), Solow (1974), Stiglitz (1974) and Hartwick (1977), introduced in the economic growth models the limitations and restrictions deriving from over use and over exploitation of natural resources.

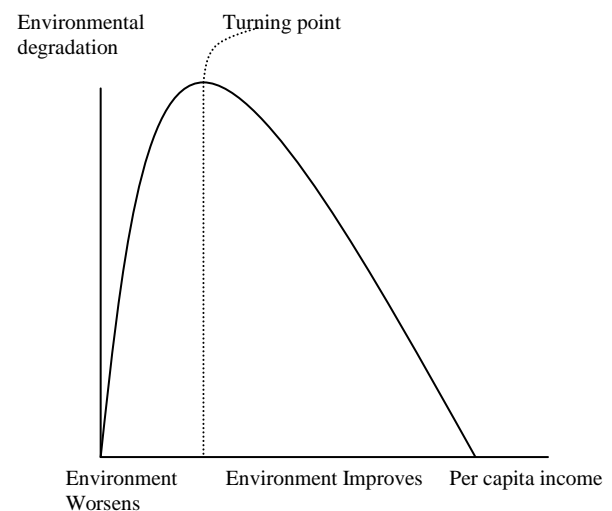
Most economic growth models that introduced the limitations of natural resources and the environment consist of a single equation specification relating an environmental impact indicator to a measure of income per capita. These models often discuss whether the relationship between income and environmental degradation actually exist, and how general and robust a model like the Environmental Kuznets Curve could be. Endogenous growth models provided the theoretical basis for income growth without complete environmental degradation or natural resource exhaustion being inevitable. Sustainability in economic growth can be achieved when the existence of resource limitations and the equity among generations are guaranteed.

Environmental Kuznets Curve

Explanations based on the relation between the environment and growth are the basis for a line of work that is known as the

Environmental Kuznets Curve (EKC). Theoretical foundations for the EKC were established by Kuznets (1955,1965,1966), which found a hump-shaped relationship between inequality and per-capita income. The implied inverted-U relationship between environmental degradation and economic growth came to be known as the Environmental Kuznets Curve, by analogy with the income-inequality relationship postulated by Kuznets (See Figure 1, below).

Figure 1. The Environmental Kuznets Curve.



An alternative explanation for the inverted relationship between certain pollutants and income per capita can be found in the hypothesized propensity of agents as they get richer to spin-off pollution-intensive products to developing countries with lower environmental standards, either through trade or direct investment in these countries (Panayotou 2000).

The EKC literature brought the empirical study of aggregate pollution levels into economic growth analysis, and changed the commonly held idea that environmental quality must necessarily decrease with economic growth (Stern, Common and Barbier 1996; Israel and Levinson 2004). The EKC literature also provides empirical evidence of a wide variety of policy response

intensities to pollution for different countries (Stern 2004).

A problem relating to the EKC curve has to do with the econometric methodology that is used for the estimation. Most of the EKC literature seeks to estimate a relationship between per capita income and pollution. Nevertheless, income and pollution are both endogenous variables that are functions of more primary determinants. Given that pollution intensities are not identical for all economic activities, it is not very likely to find a simple relationship between all possible combinations of economic growth and pollution. In fact, the shape of this relationship should vary with the source of growth, and therefore, different sources of growth will in general trace out different income and pollution paths (Bulte and Van Soest 2001).

Most of theoretical and empirical surveys on the environmental Kuznets curve are skeptical about the existence of a simple and predictable relationship between pollution and per-capita income. The EKC literature has provided empirical evidence that there is an income effect linked to environmental quality. This income effect is associated to endogenous policy responses in the form of environmental regulations that are directly related to income.

Income inequality can affect the income-environment relationship through differential marginal propensities to pollute between the rich and the poor. If the marginal propensity to pollute is higher in poor than in rich countries, then higher inequality among countries would aggregate pollution levels for any given average world income level. At the same time, any effort to improve income distribution may come at the expense of environmental quality (Ravaillon et al. 1997).

The work by Grossman and Krueger (1993) on the NAFTA can be considered to be the starting point for the literature on the

environmental Kuznets curve. The main hypothesis behind the environmental Kuznets curve is that an inverse U-shaped relationship exists between a country's per capita income and its level of environmental quality. That is, income growth could induce an increase in pollution in poor countries, but a decline in pollution in rich countries. If environmental quality is a normal good, then increases in income brought about by growth will both increase the demand for environmental quality and the ability of governments to afford costly investments in environmental protection (Copeland and Taylor 2004).

Grossman and Krueger (1993) found an inverse U-shaped between some measures of air quality and per capita income, using a panel data on air quality from 42 countries. Although some works such as Selden and Song (1994) identified a similar function using data on sulphur dioxide emissions, these results are not at all general and different relationships between growth and pollution can be found in the literature. For instance, Shafik and Banyopasdhay (1992) and Grossman and Krueger (1995) observed that pollution in contaminated drinking water declines monotonically with income per capita, while others such as carbon dioxide emissions tend to rise with income per capita. The paper by Munashinghe (1999) proposes that an environmentally adjusted measure of national income could significantly change the shape of the development-environment relationship. Using data from the United Nations Statistics Office concludes that the adoption of more sustainable policies will facilitate the attainment of higher levels of development at a lower environmental cost. Roca and others (2001), present evidence, for the case of Spain, that there is not any correlation between higher income level and smaller emissions except for SO₂ emissions, whose evolution might be congruent with the EKC hypothesis. The authors argue that the

relationship between income level and diverse types of emissions depend on many factors, and therefore it cannot be expected that economic growth, by itself will solve environmental problems. In some cases emissions could be considerably reduced by specific measures (for instance, replacing organic solvents with another type of substance). In other cases there is a bigger challenge because the reduction implies relevant changes in the present transport pattern in the sources of energy supply or in the waste management policies.

EKC and Income Effects

The income effect implies that the EKC shape reflects changes in the demand for income quality as income rises. Lopez (1994) and Copeland and Taylor (2003) demonstrate that the effects of factor accumulation on the environment depend on the interaction between the elasticity of substitution between pollution and non-polluting inputs and the income elasticity of marginal damage. Gawande and others (2001) provide a model in which agents are mobile and income effects induce a sorting equilibrium in which higher income agents avoid polluted areas.

The income effect theory can be extended to allow for political economy variables, such as corruption and institutional quality, which will move the turning point of the EKC to the right. Some empirical studies, such as Barrett and Grady (2000), include measures of political freedom as an extra shift variable in their EKC regressions. The main result is that a positive relationship seems to exist between freedom and a cleaner environment. This theory would imply that the political freedom variables interact with income variables given the expected relationship between political freedom and the policy-induced technique effect.

The income effect explanation of the EKC is based on two main hypothesis: neutral

growth and rising income elasticity of marginal damage. The neutral growth hypothesis restricts the magnitude of shifts in pollution demand as growth proceeds. The rising income elasticity of marginal damage ensures increasing technique effects. Therefore, this relationship between pollution and growth should be different for the diverse pollutants according to their perceived damage. Specifically, carbon emissions fit this model and most of studies observe that carbon emissions per capita tend to increase monotonically with per-capita income. Some works in this line are Shafik (1994), Holz-Eakin and Selden (1995). However, the work of Schmalensee, Stoker and Judson (1998) find a peak for carbon emissions per capita.

At low levels of economic activity, pollution can be unregulated entirely or regulation can have little impact on the profitability of abatement. In fact, threshold effects lead to a very different relationship between income and pollution for different stages of development (John and Pecchino 1994; Jones and Manuelli 1995).

EKC and International Trade

Empirical tests of the role of trade openness in the environment-income growth relationships date back to the origins of the EKC literature. The main hypothesis behind these models is that trade may alter environmental outcomes in very different ways. For instance, trade may encourage a relocation of polluting industries from countries with more rigid environmental policies towards those countries with more flexible or weaker regulations.

The effect of these shifts is uncertain, and could either increase global pollution or they could create weaker environmental policies because no country will be interested in more strict environmental regulations if this policy impacts on the comparative advantage of exports. Another hypothesis considered in

these models is that greater openness to trade would lead to lower environmental standards in an effort to preserve competitiveness in the face of international competition.

The work by Copeland and Taylor (1994) predicts a very different relation between growth and pollution in autarky than in free trade. If the income elasticity of marginal damage is one, then the scale and technique effects of growth exactly offset each other in autarky, and so growth has no effect on pollution. The effects created by differences in pollution regulations and in international trade rules across countries could induce these results. A main difficulty of these pool studies nowadays relies on the fact that country-specific explanations are consistent with the overall cross-country evidence. According to Dinda (2004), the reason is that the pool of observations used for cross country pool studies are based in a fundamental assumption which is that the economic development trajectory needs to be the same for all countries.

In the income effects explanation for the EKC, rich countries can reduce their pollution either by abating more or by using policy to encourage dirty industry to migrate to poor countries. If the first process is dominant, then all countries could follow a similar path. If the main driving force is the second, then even if an EKC exists for rich countries, the newly industrialized countries could not replicate the experience of the current rich (Arrow et al. 1995).

Growth and Environment

Although theoretical foundations for the environment and economic growth models can be found in the works by Ramsey (1928), as extended by Cass (1965) and Koopmans (1960), a new stream of optimal growth models including pollution or non-renewable resource depletion appeared in the 1970's closely linked to the forecasts of the Club of

Rome. These models introduced energy, natural resources and environmental pollution into the neoclassical theory of growth (Kamien and Schwarz 1982; Tahvonen and Kuuluvainen 1991,1993). Analyses following this approach have focused on the role of substitution between man-made capital and natural resource materials in production, technological improvements in materials efficiency of production, and backstop technologies (Dasgupta and Heal 1979; Van den Bergh and Nijkamp 1998). In some recent extensions of the Ramsey models a trade-off between the disutility of a pollution by-product and the utility of consumption can be found (Selden and Song, 1995; Tahvonen and Kuuluvainen, 1994).

Optimal Growth Models

In these models the social planner's problem of maximizing an infinite stream of consumption is considered and may be contrasted to the decentralized result. Models of pollution and optimal growth generally suggest that some limits to growth are optimal. On the other hand, models of natural resource depletion generally find that extraction or even extinction will be optimal depending on the discount rate and technology available in society. Some relevant models of natural resource extraction and growth are Dasgupta and Heal (1979), Stiglitz (1974), Solow (1974) and Smith (1974).

In order to include the disutility of pollution into economic activity most of the models generalize the basic dynamic optimisation of Ramsey (1928), Cass (1965) and Koopmans (1960), and yield a non-linear path like the empirically observed EKC. Empirical evidence, such as Selden and Song (1995) and Stokey (1998), suggest that these models are sensitive to assumptions about the form of the utility function and assumptions

about the insertion of pollution in the production function.

In some models the environment is introduced like a factor of production and considered a variable of the utility function. In these models environmental quality is represented as a stock that is degraded by production or pollution. In other models, the stock of environmental resources is included as a function of production and therefore the environment itself is required to generate an output (Lopez 1994; Chichilinsky 1994).

Predictions of these models regarding the effect of growth on environmental degradation depend on whether economic factors internalise the specified stock feedback effects (Panayotou 2000). The existence of the environmental stock in the production function means that optimal pollution taxes or pollution regulations are not enough to achieve the optimal level of environmental quality in the steady state.

Endogenous Growth and the Environment

The general idea in endogenous growth models is that the marginal product of human-supplied capital does not decline toward zero even as the volume of capital grows (Romer, 1986, 1990; Lucas, 1988; Barro, 1990; Rebelo, 1991). Human-supplied capital incorporates not just equipment, but also knowledge and skills. The ability to augment human as well as machine capital is one of the path ways emphasized in the theoretical assumption that the marginal product of investment can remain above some positive threshold level.

Endogenous growth models provide a theoretical approach for sustained income growth without complete environmental degradation or natural resource depletion being inevitable (Toman 2003). The models then suggest a way around limits to growth: in addition to sound natural resource and environmental practice, invest adequately in

order to build human capital. Some papers that attempt to model and endogenize technical progress within models designed to address environmental issues and sustainability are Gradus and Smulders (1993), van den Bergh and Nijkamp (1994), Bovenberg and Smulders (1995), and Smulders and De Nooij (2003). A central point in these models is that a society with strong preferences for environmental amenities could shift increasing quantities of investment toward natural capital protection as income rises, and also that a society with a high rate of discount could still choose extensive natural resource depletion.

It is important to point out that while the endogenous growth models propose a way around limits to growth, these models are strongly determined by the underlying assumptions (see the survey by Löschel 2002). A fundamental hypothesis is the ability of capital growth to generate sustained economic growth, even while flows of natural and environmental resource services remain bounded. According to Toman (2003), this seems more plausible than the simple capital-resource substitution hypothesis in the natural resource depletion models of the 1970's, but it is still not entirely self-evident.

Some models extend the ideas of endogenous growth in order to include the environment as a factor of production and environmental quality as an argument of the utility function. For instance, Bovenberg and Smulders (1995,1996) modify the model of Romer (1986) to include the environment as a factor of production. Lighthard and van del Ploeg (1986), Gradus and Smulders (1993), Stokey (1998), Smulders (1999) and Brock and Taylor (2003) extend the model used by Barro to include environmental considerations. Hung, Chang and Blackburn (1994) use the Romer (1990) model, and Sala and Subramanian (2003) show that some natural resources, (oil and minerals in

particular), exert a negative and nonlinear impact on growth via their deleterious impact on institutional quality.

The endogenous growth models augmented in order to include environmental changes reinforce the results of neoclassical growth models with respect to environmental degradation. Although models are sensitive to the specification of the utility function, results indicate that optimal pollution control requires a lower level of growth than would be achieved in the absence of pollution.

The endogeneity of pollution policy is a central assumption in the theoretical and empirical literature. If governments are sensitive to pollution, then pollution regulations could become more inflexible and rigid and positive income growth should lead to an increase in the demand for environmental quality.

Redistribution may interact with growth and the trade off between growth and environmental quality could improve with redistribution. Another way in which inequality may affect environmental outcomes is by strengthening the power of the rich to impose environmental costs on the poor (Boyce 1994) and by reducing the ability of the society to reach cooperative solutions to environmental problems (Ostrom 1990).

Vincent (1997a) found that the net impact of population density on total suspended particulates concentration was positive because household activities are important sources of particulate concentrations. Vincent (1997b) also found a negative interaction term between population density and time, indicating a downward pressure on population-driven total suspended particulates concentrations by the simply passage of time, which he attributes to increasingly effective anti-pollution regulations. The same results were found for water quality.

Kauffman et al. (1998) considered the inclusion of population density in a multiplicative relationship with per capita GDP, in order to obtain the spatial intensity of economic activity. The main intuition is that increasing population density is likely to have a small effect on SO₂ concentrations when per capita GDP is low, and therefore, emissions per person are low. Increasing population density has a much larger effect when per capita GDP is high and emissions per person are high. Kauffman et al. (1998) specified a reduced form equation for SO₂ concentrations, quadratic in both GDP per capita and economic activity per unit of area, which they tested with panel data for 23 developed, developing and transitional economics for the years 1974-1989. The authors found a U-shaped relationship between atmospheric concentrations of SO₂. Concentrations tend to decrease as GDP per capita rises from \$3000 to \$12500 and increase thereafter. These changes are attributed to the energy use that is linked to economic development, with a shift towards cleaner fuels dominating below \$12500 and the increase in energy consumption dominating at higher incomes.

The income–environment relationship specified and fitted in the literature aims to calculate the net effect of income on the environment. Three main structural forces have been characterized: (Panayotou 1997; Kauffman et al. 1998; Islam, Vincent and Panayotou 1999; Nguyen 1999):

1. The scale of economic activity.
2. The structure of the economic activity.
3. The effect of income on the demand and supply of pollution abatement efforts.

The scale effect on pollution is expected to be a monotonically increasing function of income, given that the larger the scale of economic activity per unit of area, the higher the level of pollution all else equal. The

structural change that accompanies economic growth affects environmental quality by changing the composition of economic activity toward sectors of different pollution intensity. At lower levels of income, the dominant shift is from agriculture to industry, and hence there will be an increase of pollution intensity. If the dominant change is for industry to service, then there will be a decrease in the pollution intensity. The composition effect is then likely to be a non-monotonic (inverted-U) function of GDP. Once the scale and composition effects of income growth are taken into account, then pollution is a non-increasing function of income, reflecting the non-negative elasticity for environmental quality.

Changes in the income-environment relationship in the course of economic growth have been attributed both to structural and behavioural factors (Panayatou 2000). It was in the 1970's that variables such as energy, natural resources and environmental pollution were introduced in the neoclassical theory of growth. The Club of Rome predictions are crucial in explaining this change in the theoretical foundations of the model. Twenty years later, in the 1990's, the report of the Brundland Commission (World Commission on Environment and Development 1987), introduced these considerations into the endogenous growth theory models. The main difference between endogenous growth models and neoclassical models is that the stationary growth rate in endogenous growth models can be positive even if there is no hypothesis about some variable growing exogenously (for instance technology or natural resources in the neoclassical models). The stationary growth rate is determined by endogenous variables rather than exogenously set variables.

Jaffe, Newell and Stavins (2002) emphasize the relevance of technical change in the context of environmental policy. It is

difficult to reject that the effects of environmental policies on the development and spread of new technologies can be, in the long run, among the most important determinants of success or failure of environmental protection efforts (Kneese and Schultze 1975). At the same time, it has long been recognized that alternative types of environmental policy instruments can have significantly different effects on the rate and direction of technological change (Orr 1976). As suggested by Kemp and Soete (1990), environmental policies, particularly those with large economic impacts can be designed to foster rather than inhibit technological invention, innovation and diffusion.

Policies linking environmental policy instruments and technological change can be characterized as either command-and-control or market-based approaches (Jaffe, Newell et al 2002). Command-and-control regulations are usually considered to allow relatively little flexibility in the means of achieving technological goals. This policy tools tend to force firms to take uniform standards, mostly technology and performance-based standards. Technology-based standards specify the method and sometimes the equipment that firms should use to comply with a particular regulation. A performance standard sets a uniform control target for firms, while allowing some latitude in how this target should be achieved. The objective of Market based policies is to encourage firm behaviour through market signals rather than with the use of explicit regulation.

Economic Growth and Sustainability

An economy is deemed to be sustainable, (in a weak sense), if the ratio of savings to income, (which allows investment), is larger than the sum of the ratios of depreciation of human-made capital and "natural capital" (Martínez-Alier 1995). Economic growth will be weakly sustainable whenever the ratio of

savings to income is larger than the sum of the ratios of depreciation of human-made capital and the stock of natural capital. Pearce and Atkinson (1993) have defined sustainability in a strong sense by maintaining a critical natural capital constant. Pearce and Atkinson (1993) have defined sustainability also in a “strong sense”, which implies maintaining critical “natural” capital constant. The Brundtland Commission defined the term sustainable development as “development that meets the needs of present generations without compromising the ability of future generations to meet their needs”. This definition calls for attention to intergenerational justice with respect to the use of the world’s limited resources. It can be explained in terms of natural capital. If the stock of natural capital exceeds the sustainability needs, then present and future generations could increase their consumption and use of natural resources. Obviously, this situation is inefficient. On the other side, if the current natural stock is smaller than the desired sustainable stock, then the current generation should be forced to save and reduce the use of natural resources if sustainable steady state consumption per capita should have to be achieved. This decision will depend on the preferences of the current generation.

The two main factors common to most definitions of sustainable growth are the existence of resource limitations and equity among generations (Chichilnisky, Heal and Beltritti 1995). According to Aghion and Howitt (1998), growth will be sustainable in the face of environmental pollution when the elasticity of intertemporal substitution in consumption is less than one. This implies that people is rational enough to avoid a critical threshold level of environmental quality below which the environment would be likely to disappear.

According to Solow (1993), it makes no sense to interpret sustainable development as requiring generations to leave each and every resource stock in its initial situation. That is, there are ways of substituting one kind of resource for another (nevertheless, this substitution could imply a reduction in the number of species available for future generations). Other researchers, such as Pearce, Barbier and Markandya (1990) have argued that although some substitution is possible, each generation should leave intact the overall stock of natural capital for which other kinds of capital cannot be a substitute.

This implies that although it is true that the degree of sustainability between various kinds of capital is often low and hard to estimate with precision, nevertheless any realistic assessment of the demands placed by the current environment users by intergenerational equity must define those demands not in terms of moral obligations to preserve this but in terms of an obligation to leave an adequate capacity for material development, where that capacity is to be represented by a comprehensive measure of capital.

Given that, one kind of capital can be substituted for another, so there is a trade-off that can be exploited between the material enjoyment of manufactured commodities and the preservation of raw materials. This implies that not allowing any fossil fuel to be used would be to deprive humanity of much of the material benefits that technological progress has made possible over the generations. To allow it to be used at no cost at all in order to maximize the current flow of goods would be also a wrong decision. Some trade off must be sought and therefore welfare should have to be defined in a way that could recognize some substitutability between manufactured goods and ecological or environmental concerns. Some authors like Ropke (2001), point out that most

technological changes are motivated by completely different reasons than environmental considerations and that a very important way to renew consumer goods is based in the use of new core technologies.

Endogenous growth theory is considered to be more suitable for addressing the problems of sustainable development than neoclassical theory, because a central concern of the endogenous growth theory is whether or not growth can be sustained (Aghion and Howitt, 1998). One of the most important conclusions of endogenous growth models related to the environment is that the chances of achieving sustainable growth depend critically on maintaining a steady flow of technological innovations. In these models, not only the capital is considered, (as in the more aggregated approach), but a distinction is introduced between innovation and capital accumulation, in order to capture the critical role of innovation in making growth sustainable. More specifically, it turns out to be crucial for sustainability in the face of natural resource constraints that the technology for producing knowledge is generally cleaner than that for producing physical capital.

Conclusions

The main interest of discussing the relationship between Economic Growth and the Environment is the evaluation of the economic and social aspects of economic policies and its sustainability. It is still hard to derive strong conclusions from environmental economic growth models, mainly because of the difficulties to price and to assign some precise economic values to the environmental goods that compose the natural capital. The Environmental Kuznets Curve is a powerful economic tool in order to enhance the analysis of the effects deriving from economic and ecological policies.

Endogenous growth models are effective to clarify some of the issues raised by the concept of sustainable development. Mainly because these models allow the introduction of relevant variables, (such as pollution, energy or the stock of natural resources), and because they are a very useful tool to improve the accuracy in the characterization of sustainable development.

Endogenous growth models are a good methodological scenario for the interdisciplinary analysis in which we can find that causality between economic growth and the environment does not go only in one sense, and that changes in one variable could present effects on the others, no matter where the first causation comes from.

Improvements in the analysis can be expected from an enhancement in the quality and in the amount of data. New data could allow clarification of theoretical models used to explain economic growth, mainly from the explicit consideration of the interrelationship between economics and environmental sciences. From these interdisciplinary models, a more meticulous definition of sustainability can be expected, and therefore a more accurate calculation and estimation of economic growth models and its economic, social and environmental effects.

Selected References

- Aghion, P. and P. Howitt (1998) *Endogenous Growth Theory*. MIT Press, Cambridge.
- Arrow, K.; B. Bolin; R. Constanza; P. Dasgupta; C. Folke; C.S. Holling; B.O. Janson; S. Levin; K.G. Maler and C. Perrings. (1995) "Economic Growth, Carrying Capacity and the Environment", *Science*, 268, 520-521.
- Barrett, S. and K. Grady. (2000) "Freedom, Growth and the Environment", *Environment and Development Economics* 5, 4, 433-456.

- Barro, R.J. (1990) "Government Spending in a Simple Model of Endogenous Growth", *Journal of Political Economy*, 98, 5, Part 2, 103-125.
- Bovenberg, A.I., and S. Smulders. (1995) "Environmental Quality and Pollution-Augmenting Technological Change in a Two-Sector Endogenous Growth Model", *Journal of Public Economics*, 57, 369-391.
- Bovenberg, A.I., and S. Smulders. (1996) "Transitional Impacts of Environmental Policy in an Endogenous Growth Model", *International Economic Review*, 37, 861-893.
- Boyce, J.K. (1994) "Inequality as a Cause of Environmental Degradation", *Ecological Economics*, 11, 169-178.
- Brock, W.A. and M.S. Taylor. (2003) The Kindergarten Rule of Sustainable Growth. NBER Working Paper 9597. New York: NBER.
- Bulte, E.H., and D.P. Van Soest (2001) "Environmental Degradation in Developing Countries: Households and the (Reverse) Environmental Kuznets Curve", *Journal of Development Economics*, 65, 1, 225-235.
- Cass, D. (1965) "Optimum Growth in an Aggregative Model of Capital Accumulation", *Review of Economic Studies*, 32, 233-240.
- Chichilinsky, G. (1994) "Global Environment and North-South Trade", *American Economic Review*, 84, 4, 851-874.
- Chichilnisky, G.; G. Heal and A. Beltritti (1995) "The Green Golden Rule", *Economics Letters*, 49, 175-179.
- Copeland, B.R., and M.S. Taylor. (2004) "Trade, Growth and the Environment", *Journal of Economic Literature*, 42, 7-71.
- Copeland, B.R., and M.S. Taylor (2003) *Trade and the Environment: Theory and Evidence*. Princeton, NJ: Princeton University Press.
- Copeland, B.R., and M.S. Taylor (1994) "North-South Trade and the Environment", *Quarterly Journal of Economics*, 109, 3, 755-787.
- Dasgupta, P.S. and G.M. Heal. (1974) "The Optimal Depletion of Exhaustible Resources", *Review of Economic Studies*, 41, 3-28.
- Dasgupta, P. S. and G.M. Heal. (1979) *Economic Theory and Exhaustible Resources*. Cambridge: Cambridge University Press.
- Dinda, S. (2004) "Environmental Kuznets Curve Hypothesis: A Survey", *Ecological Economics*, 49, 4, 431-455.
- Edward-Jones, G.; B. Davies and S. Hussain (2000) *Ecological Economics*. London: Blackwell Science.
- Gawande, K.; R. Berrens and A.K. Bohara (2001) "A Consumption Based Theory of the Environmental Kuznets Curve", *Ecological Economics*, 37, 1, 101-112.
- Gradus, R. and S. Smulders. (1993) "The Trade-Off between Environmental Care and Long Term Growth Pollution in Three Prototype Growth Models", *Journal of Economics—Zeitschrift fuer Nationaloekonomie*, 58, 25-51.
- Grossman, G. and A. Krueger (1993) "Environmental Impacts of a North American Free Trade Agreement", in Peter Garber (Editor), *The US-Mexico Free Trade Agreement*. Cambridge, MA: MIT Press.
- Grossman, G.M., and A.B. Krueger. (1995) "Economic Growth and the Environment", *Quarterly Journal of Economics*, 110, 2, 353-377.
- Hartwick, J.M. (1977) "Intergenerational Equity and the Investing of Rents from Exhaustible Resources", *American Economic Review*, 67, 5, 972-974.
- Holz-Eakin, D. and T. Selden (1995) "Stoking the Fires? CO₂ Emissions",

- Journal of Public Economics*, 57, 1, 85-101.
- Hung, V.T.Y.; P. Chang and K. Blackburn (1994) "Endogenous Growth, Environment and R&D", in C. Carraro (Editor), *Trade, Innovation and Environment*. The FEEM/KLUWER International Series on Economics, Energy and Environment (Economics, Energy and Environment). London: Springer.
- Islam, N.; J.R. Vincent and T. Panayotou (1997) *Unveiling the Income-Environment Relationship: An Exploration into the Determinants of Environmental Quality*. Boston: Harvard Institute for International Development, Development Discussion Paper 701.
- Israel, D. and A. Levinson (2004) "Willingness to Pay for Environmental Quality: Testable Empirical Implications of the Growth and Environment Literature", *Contributions to Economic Analysis and Policy*, 3,1, 1254ff. Berkeley: Berkeley Electronic Press.
- Jaffe, A.B.; R.G. Newell and R.N. Stavins (2002) "Technological Change and the Environment", in K.G. Maler and J. Vincent (Editors), *Handbook Environmental Economics*. Volume I. Amsterdam: North Holland/Elsevier.
- John, A. and R. Pecchino (1994) "An Overlapping Generations Model of Growth and the Environment", *Economic Journal*, 104, 1393-1410.
- Jones, L. and R. Manuelli (1995) *A Positive Model of Growth and Pollution Controls*. NBER Working Paper 5205. New York: NBER.
- Kauffman, J.B.; D.L. Cummings and D.E. Ward (1998) "Fire in the Brazilian Amazon. 2. Biomass, Nutrient Pools and Losses in Cattle Pastures", *Oecologia*, 113, 415-427.
- Kemp, R. and L. Soete. (1990) "Inside the 'Green Box': On the Economics of Technological Change and the Environment", in C. Freeman and L. Soete (Editors), *New Explorations in the Economics of Technological Change*. London: Pinter.
- Kneese, A. and C. Schultze (1975) *Pollution, Prices and Public Policy*. Washington, D.C.: Brookings Institution.
- Koopmans, T.C. (1960) "Stationary Ordinal Utility and Impatience", *Econometrica*, 28, 287-309.
- Kuznets, S. (1965) *Economic Growth and Structural Change*. New York: Norton.
- Kuznets, S. (1966) *Modern Economic Growth*. New Haven: Yale University Press.
- Kuznets, S. (1955) "Economic Growth and Income Inequality", *American Economic Review*, 45, 1, 1-28.
- Lopez, R. (1994) "The Environment as a Factor of Production: The Effects of Economic Growth and Trade Liberalization", *Journal of Environmental Economics and Management*, 40, 2, 137-150.
- Löschel, A. (2002) *Technological Change in Economic Models of Environmental Policy: A Survey*. FEEM Working Paper 4.2002. Milan, Italy: Fondazione Eni Enrico Mattei.
- Lucas, R. (1988) "On the Mechanics of Economic Development", *Journal of Monetary Economics*, 22, 1, 3-42.
- Martínez-Alier, J. (1995) "The Environment as a Luxury Good or Too Poor to Be Green", *Ecological Economics*, 13, 1-10.
- Meadows, D.H.; D.L. Meadows; J. Randers and W. Behrens (1972) *The Limits to Growth*. Universe Books, New York.
- Munashinghe, M. (1999) "Making Economic Growth More Sustainable", *Ecological Economics*, 15, 121-124.
- Nguyen, A.T. (1999) *Evidences of the Environmental Kuznets Curve from CO₂ Emissions in Six Country Analyses*.

- Working Paper. Grenoble, France: Département Energie et Politiques de l'Environnement, Université Pierre Mendès.
- Orr, L. (1976) "Incentive for Innovation as the Basis for Effluent Charge Strategy", *American Economic Review*, 66, 441-447.
- Ostrom, E. (1990) *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, Cambridge.
- Panayotou, T. (1997) "Demystifying the Environmental Kuznets Curve: Turning a Black Box into a Policy Tool", *Environmental and Development Economics*, 2, 4, 465-484.
- Panayotou, T. (2000) "Economic Growth and the Environment", CID Working Paper 56, *Environment and Development Paper*, 4.
- Pearce, D.; E. Barbier and A. Markandya (1990) *Blueprint for a Green Economy*. Earthscan Publications, London.
- Pearce, D. and G.D. Atkinson (1993) "Capital Theory and the Measurement of Sustainable Development: an Indicator of Weak Sustainability", *Ecological Economics*, 8, 103-108.
- Ramsey, F. (1928) "A Mathematical Theory of Saving", *Economic Journal*, 38, 543-559.
- Ravaillon, M.H., and J. Jalan (1997) *A Less Poor World, but a Hotter One? Carbon Emissions, Economic Growth and Income Inequality*. Washington DC: World Bank.
- Rebelo, S. (1991) "Long Run Policy Analysis and Long Run Growth", *Journal of Political Economy*, 99, 3, 500-521.
- Roca, J., Padilla, E., Farre, M., and V. Galletto (2001) "Economic Growth and Atmospheric Pollution in Spain: Discussing the Environmental Kuznets Curve Hypothesis", *Ecological Economics*, 39, 85-99.
- Romer, P. M. (1986) "Increasing Returns and Long Run Growth", *Journal of Political Economy*, 94, 5, 1002-1037.
- Romer, P. M. (1990) "Endogenous Technical Change", *Journal of Political Economy*, 98, 5, Part 2, 71-102.
- Ropke, I. (2001) "The Environmental Impact of Changing Consumption Patterns: a Survey", *International Journal of Environment and Pollution*, 15, 2, 127-145.
- Sala, X., and A. Subramanian. (2003) *Addressing the Natural Resource Curse: An Illustration from Nigeria*. Working Paper 03/139. Washington DC: IMF.
- Schmalensee, R.; T. Stoker and R. Judson (1998) "World Carbon Emissions", *Review of Economics and Statistics*, 80, 1, 15-27.
- Selden, T. and D. Song (1994) "Environmental Quality and Development: Is There a Kuznets Curve for Air Pollution Emissions?", *Journal of Environmental Economics and Management*, 27.2, 147-162.
- Selden, T.M., and D. Song (1995) "Neoclassical Growth, the J Curve for Abatement and the Inverted U Curve for Pollution", *Journal of Environmental Economics and Management*, 27.2, 147-162.
- Shafik, N. (1994) "Economic Development and Environmental Quality: An Econometric Analysis", *Oxford Economic Papers*, 46, 757-773.
- Shafik, N. and S. Banyopasdhay (1992) *Economic Growth and Environmental Quality: Time Series and Cross Sectional Evidence*. Working Paper 904. Washington DC: World Bank.
- Smith, V. L. (1974) "An Optimistic Theory of Exhaustible Resources", *Journal of Economic Theory*, 9, 384-396.
- Smulders, S. (1999) "Endogenous Growth Theory and the Environment", in J.C.J.M. Van den Berg (Editor), *Handbook of Environmental Resource Economics*, Edward Elgar, Cheltenham, 610-621.

- Smulders, S. and M. De Nooj (2003) "The Impact of Energy Conservation on Technology and Economic Growth", *Resources and Energy Economics*, 25, 59-79.
- Solow, R.M. (1974) "Intergenerational Equity and Exhaustible Resources", *Review of Economic Studies*, 41, 29-45.
- Solow, R.M. (1993) "Sustainability: An Economist's Perspective", in R. Dorfman and S. Dorfman (Editors), *Economics of the Environment: Selected Readings*. Third Edition. Norton, New York.
- Stern, D.I. (2004) "Comments on: Cole, M. A. (2003) Development, Trade and the Environment: How is the Environmental Kuznets Curve?", *Environment and Development Economics*, 8, 557-580.
- Stern, D.I; M.S. Common and E.B. Barbier (1996) "Economic Growth and Environmental Degradation: The Environmental Kuznets Curve and Sustainable Development", *World Development*, 24, 1151-1160.
- Stiglitz, J.E. (1974) "Growth with Exhaustible Natural Resources", *Review of Economic Studies*, 41, 139-152.
- Stokey, N. (1998) "Are There Limits to Growth?", *International Economic Review* 39,1, 1-31.
- Tahvonen, O. and J. Kuuluvainen. (1994) "Economic Growth, Pollution and Renewable Resources", *Journal of Environmental Economics and Management*, 24, 2, 101-118.
- Toman, M. (2003) *The Roles of the Environment and Natural Resources and Natural Resources in Economic Growth Analysis*. Discussion Paper 02-71, *Resources for the Future*.
- Vincent, J. (1997a) "Pollution and Economic Development in Natural Resources, Environment and Development", in Jeffrey R. Vincent, Rozali Mohamed Ali, and Associates (Editors), *Environment and Development in a Resource-Rich Economy: Malaysia under the New Economic Policy*, Boston: HIID Books.
- Vincent, J. (1997b) "Testing for Environmental Kuznets Curves Within a Developing Country", *Environment and Development Economics*, 2, Part 4, 417-433.
- World Commission on Environment and Development (1987) *Our Common Future*. Oxford: Oxford University Press.

Webites

Economic Growth Center:

www.econ.yale.edu/~egcenter

Economic Growth Resources:

www.bris.ac.uk/Depts/Economics/Growth

World Bank. Macroeconomics & Growth:

www.econ.worldbank.org

Oscar Alfranca

Agrifood Engineering & Biotechnology Dept

Universitat Politècnica de Catalunya

Madrid, Spain

oscar.alfranca@upc.es

Education Policy: Distance

James J.F. Forest

Introduction

Recent decades have seen a dramatic worldwide expansion in distance education courses and programs, covering the entire spectrum of academic disciplines and subject areas. The term distance education is generally used to describe instructional activities where the teachers and learners are separated by physical space (and often time). Despite this separation, expectations of teaching and learning are much the same as in traditional classroom instruction. Teachers are responsible for providing knowledge and guidance in the subject matter, and assessing their students' development. Students are responsible for studying the assigned learning materials and demonstrating their achievement of subject proficiency through exams, papers, projects, and other assignments. The learning objectives of courses offered through classroom instruction or distance education are usually comparable. It is widely believed that student outcomes of a course should be the same regardless of the way in which these objectives are met.

From a public policy perspective, the potential benefits of distance education are striking, particularly in terms of meeting the increasing demands for access to higher education. As newer and increasingly affordable technologies have enabled the delivery of educational services to a wider range of students, national policymakers in many parts of the world have targeted their public financial investments toward distance education programs. Examples include the University of the West Indies, China's Central Radio and Television University, the DIRECWAY Global Education initiative in India, Canada's Athabasca University, the Mauritius College of the Air, Germany's

FernUniversität, and Japan's University of Air. Throughout Africa, a variety of correspondence colleges were established during the 1960s and 70s, including in Botswana, Malawi, Tanzania, and Zambia. One of the most significant initiatives on the continent has been the work of INADES (*Institut Africain pour le Développement Economique et Social*) which from its headquarters in Abidjan has since 1962 been providing correspondence courses to several francophone and anglophone countries in agriculture and agricultural economics (Jenkins, 1989). Today, the forces of globalization, as well as inter-governmental organizations like the World Bank, International Monetary Fund and UNESCO, are aiding in the global spread of distance education, guided by the principles of human capital and neo-liberal development theories.

Early forms of distance education involved correspondence learning, where interaction between the student and instructor took place through the mail (Blanke and Wekke 2002). As new forms of technology were developed (such as photos, telephone, audio and video recordings, broadcast radio and television) these were used to enhance the learning experience for students in distance education programs. Since the 1960s, instructional television has made a particularly significant impact on the delivery of recorded and live instruction to a wide audience. More recently, satellites, videoconferencing, multimedia closed circuit networks and the Internet have further expanded the reach and attraction of distance education (Blanke and Wekke 2002). By combining text, graphics, sound and video, multimedia technologies provide a richness in educational content previously unavailable in other distance education media, while new communications technologies allow for greater levels of interaction between instructor and student, and among students,

which many believe improves upon more passive media such as television and video (Snydman 2002).

A brief review of today's higher education landscape reveals a wide range of distance education activities in which millions of students worldwide participate annually. In some cases, distance education is used to supplement regularly scheduled classroom-based courses, while others provide pure distance learning courses that require no face-to-face contact between instructors and students. Some institutions may offer distance education courses in addition to a college's regular course listings, or sometimes a distance learning course is used to replace a previously classroom-based course (Petrides 2002). In many parts of the world, entire postsecondary and graduate degree programs are available via distance education, while a growing number of "virtual universities" (which have no physical classroom space, and instead support all teaching and learning primarily via the Internet using Web-based technologies) are offering both opportunities for students and challenges for a nation's public universities and policymakers.

There are typically four types of public policy debates common to any nation currently exploring initiatives to support (or constrict) the delivery of distance education. The first is the challenge of providing access to distance education in largely rural, underdeveloped regions of a country. The second is the costs associated with infrastructure needs of distance education. The third is the decision of whether or not to allow for-profit institutions (local or foreign) to offer distance education programs. And the fourth is assessing and ensuring the quality of distance education courses and programs.

Access to Higher Learning

The most prominent public policy implication of distance education has been the promise it

offers for meeting the increasing demand for access to higher education. In many countries, a large segment of the population is unable to attend traditional residential or commuter institutions of higher education for a variety of reasons, including family responsibilities, physical disabilities, and geographic proximity to a college or university. Particularly in developing countries, the opportunity to enroll in a higher education institution is typically limited to individuals who live in or near the major urban areas. The justification for investing in distance education initiatives has thus been quite strong in these countries, particularly in terms of programs that provide largely rural populations with the knowledge and skills to improve their agricultural productivity.

Today, distance learning initiatives play a vital role in providing access to higher education. From extension courses to full-scale degree programs and online institutions, distance education opportunities are particularly common throughout Asia and the Pacific (with 15 digital universities in Korea and 67 online colleges established by conventional universities in China), Latin America (particularly Brazil, where 80,000 students were enrolled in distance education courses in the 2000-01 academic year), Africa and the Arab World. New institutions in Africa include Zimbabwe Open University, the Open University of Tanzania, and the African Virtual University, all of which were established after 1997. The National Open University of Nigeria, which was closed in 1985, re-opened its doors in May 2003 with 100,000 students attending 18 campuses distributed throughout the country. The Arab Open University was established in 1999 with its headquarters in Kuwait and branches in Bahrain, Egypt, Lebanon, Jordan and Saudi Arabia. The Syrian Virtual University, established in 2002, offers internationally accredited degrees, and has recently

concluded agreements with western online universities (particularly in Canada and the United States) to provide their programs to Arab students.

Despite these developments, critical challenges exist—particularly for rural communities, where the infrastructure necessary for providing distance education is severely limited. The lack of Internet connectivity, phone service, or even reliable electricity severely undermines the development of viable distance education programs in these regions. The public policy implications here are striking, given that many rural populations (indeed, nearly one fifth of humanity) are functionally illiterate and are thus limited in their capacity for productive participation in society. Providing basic distance education to rural areas has thus become a major public policy concern throughout the world. Overall, the evolution of technology has changed the landscape of higher education and, by implication, the social and economic benefits a country can derive from investments in distance education initiatives. However, while meeting access is a key impetus behind the tendency of many countries to embrace distance education, policymakers are beginning to raise concerns over how to manage, measure, and understand the costs associated with these initiatives.

Managing the Costs of Initiatives

Obviously, with limited public funds (particularly in developing countries) policymakers are already under considerable pressure to ensure their spending decisions yield the largest return on the public's investment. Thus, they are understandably alarmed to find that the costs of developing the infrastructure necessary to support a robust distance education experience can be staggering, and budget plans often fail to account for the ongoing costs of maintenance,

upgrades and support (Snydman 2002). Further, policymakers must consider a number of important risks that are inherent in making infrastructure investments to support distance learning.

For example, national leaders and higher education administrators looking to develop distance education programs find themselves having to make decisions about Internet-related technologies in an environment of fast-paced change, obsolescence, and uncertainty (Petrides 2002). At any moment in time, choosing a particular configuration of software, hardware, network infrastructure, and information technology support staff (often thru million-dollar contracts) requires a great deal of trust that significantly better products will not come along anytime soon, and that the makers of the products purchased will remain in business and provide some form of service to help keep the distance education program running. Making a wrong decision at this juncture could have disastrous effects both for the distance education initiative as well as the public's confidence in their country's leaders. Similar risks are seen at the college or university level as well. The costs of providing distance education are often prohibitively expensive for the budgets of higher education institutions, and making a bad technological investment can be disastrous to the long-term future of the institution.

There are, of course, a variety of financial benefits from investing (wisely) in distance education. It is certainly the case that the one-to-many nature of Internet technologies allows courses to be delivered to large audiences with much smaller investments in faculty and staff (Snydman 2002). A single faculty member can (with appropriate technical and logistical support) instruct thousands of students in a single semester. The lectures of prominent experts in a particular academic subject or specialty can

be recorded on audio and video and transmitted repeatedly to various audiences at any time of day or night. With those distance education programs that utilize e-mail or web-chat applications, the level of communication between student and teacher can actually be greater than what is available in a classroom (and particularly large lecture halls). However, amid these benefits are a number of embedded costs that must be managed carefully and wisely.

Private and For-Profit Providers

Another important public policy debate surrounding distance education involves the increasing role of for-profit and private institutions in the provision of higher education. Indeed, private initiatives have made significant headway in the realm of distance education, offering both short-term and semester-long courses as well as entire degree programs. For example, the Open University (based in the United Kingdom) and Jones International University (based in the United States) have both established a truly global presence, with programs offered in dozens of countries. The accreditation of such programs and “virtual universities” in recent years heralds a new era for distance education. These private institutions and (increasingly) for-profit companies have responded to the need for large-scale applications used for distance learning by launching their own applications for commercial use as well as their own online courses or virtual universities (Petrides 2002).

Most of the private and non-profit institutions offering distance education are based in North America and Europe. Thus, leaders of countries throughout the rest of the world grapple with the decision of whether to allow these institutions to operate within their borders. On the one hand, for a developing country to allow a foreign higher education provider into the country appears to some as

an admission to the inadequacy of the country’s ability to meet on its own the demand for higher education. In addition, these foreign providers are competing with local institutions for fee-paying students, thus potentially undermining the ability of local colleges and universities to survive. On the other hand, if local institutions are seen as lacking in quality and access, and if public funds are inadequate to support indigenous distance education programs, a country’s leaders may see no other choice but to invite foreign providers in. In either case, decisions regarding foreign private and non-profit providers of distance education are bound to be unpopular with at least some segments of the local population. And the increasing globalization of higher education will continue to exacerbate these tensions for many years to come.

Quality Assurance

Because of the need to provide greater access to higher education, while managing costs and avoiding an over-reliance on the private sector, public policymakers throughout the world are showing an increasing interest in measuring the outcomes of university activities. Indeed, ‘return on investment’ is a term one commonly hears in discussions on the public funding of higher education, and in some countries, funding decisions have become closely tied with assessments of performance quality. In distance education, questions regarding quality tend to focus on how effective communication technologies are as substitutes for face-to-face interaction between the student and the instructor (Snydman 2002). Indeed, one of the most heated topics of debate in both academic and public policy circles concerns the issue of whether distance education provides a learning experience for students that is comparable to that of traditional “bricks and mortar” institutions.

Current research in this area is mixed. Some studies have suggested that students enrolled in distance education programs demonstrate a familiarity with subject matter that is fairly equal to their counterparts enrolled in traditional colleges and universities. Similar research has suggested there is virtually no difference in terms of objective examination results between distance education students and traditional students (cf. Phipps and Merisotis 1999). Meanwhile, critics of distance education question the rigor of distance education assignments and examinations, and argue that there can be no valid substitute for traditional classroom instruction. Others contend that while practices such as monitoring correspondence teaching, visiting tutors in study centers, and assessing technical infrastructure have been common in many distance education programs through the years, these efforts have represented less *quality assurance* and more *quality control* (Tait 1997).

Within this sometimes heated debate, it falls to public policymakers to determine the criteria and methods by which distance education programs should be evaluated. In some cases, distance education providers are subjected to the same accreditation reviews as their traditional institutional counterparts. Examples of national agencies responsible for these evaluations include the Australian Universities Quality Agency, Thailand's Office of Education Standards and Evaluation, Argentina's National Commission for Evaluation and Accreditation (CONEAU), Brazil's National Education Council and Higher Education Board, Chile's National Commission of Undergraduate Accreditation (CNAP), and Costa Rica's National Accreditation System for Higher Education.

In other cases, regional consortia have been (or are in the process of being) formed

to develop and implement a standardized approach to the evaluation of distance education providers. For example, the South African Development Community (SADC) and the Economic Community of West African States (ECOWAS) are serving a vital role in identifying common assessment indicators and standards for the recognition of studies and degrees in both the traditional and distance education arenas. And inter-governmental and non-governmental organizations are playing an increasingly visible role in promoting the development and assessment of distance education. Examples include the UNESCO-CEPES project *Strategic Indicators for Higher Education in the 21st Century* and the International Network of Quality Assurance Agencies in Higher Education (INQAAHE).

Conclusion

Subsumed by these four public policy debates is a necessary focus on the learner in distance education. To a much greater degree than in traditional higher education, students have a tremendous responsibility for their learning in any distance education activity. Whether it be a correspondence college or a "real-time" two-way communication activity via the Internet, students must meet deadlines for reading and writing assignments, while taking a more active role in making sure the learning experiences provided by the distance education program meet their needs and expectations. Thus, in some ways distance education initiatives are causing instructors to approach their teaching with a new mindset, one in which the learning is seen as more of a collaborative venture than the traditional "pouring knowledge into the empty vessel" concept. While this can certainly be a good thing for both the student and teacher, it also raises expectations of "customer satisfaction", which has implications for public perceptions of an institution's quality and success—and,

by extension, the success of distance education-related public policies.

Further, this consumer-related shift of perception—with its focus on customer satisfaction—leads directly to changes in how a country's citizens and leaders view the provision of higher education. In essence, higher education is less likely to be seen as a public good, but more likely to be considered a private good. Here, the public policy implications of distance education become particularly salient when considering the need for any country's leaders to ensure broad public support for their decisions. If distance education programs are seen as a private good, the nation's policymakers will find less support for using public funds to develop and maintain them. This forces distance education providers (both public and private) to shift the burden of costs to students and their families, which by extension has an impact on the level and types of access to higher education available.

In sum, the phenomenon of distance education as a whole can be said to have at least some marginal impact on the relationship between individuals and their elected leaders, particularly in the realm of providing access to quality education. How this shifting relationship is managed is clearly something to watch closely in the coming decades, particularly in large, developing countries where distance education has become an increasingly attractive vehicle for providing access to higher education. At the very least, as long as distance education programs are viewed by the public as meeting a vital need for skills development and credentials, and as long as the student outcomes are seen as having an acceptable level of quality, these programs will continue to flourish.

Selected References

- Bates, Anthony W. (1995) *Technology, Open Learning, and Distance Education*. London: Routledge.
- Blanke, Debra J. and Gina M. Wekke. (2002) Distance Education. In James J.F. Forest and Kevin Kinser (Editors), *Higher Education in the United States: An Encyclopedia*. 2 volumes. New York: ABC-CLIO Publishers, 174-177.
- Butcher, Neil and Nicky Roberts. (2004) "Costs, Effectiveness, Efficiency", in Hilary Perraton and Helen Lentell (Editors), *Policy for Open and Distance Learning*. London: RoutledgeFalmer.
- Chronicle of Higher Education*. (1999) "The Marketing Intensifies in Distance Learning", 9 April.
- Chronicle of Higher Education*. (1999) "Virtual Universities Can Meet High Standards", 4 November.
- Fergus, Howard A. (1998) "From Experiment to Enterprise: Distance Teaching at the University of the West Indies", in James JF Forest (Editor), *University Teaching: International Perspectives*. New York: Garland Publishing, Inc., 345-360
- Finkelstein, Martin, Claire Frances and Bruce W. Scholz. (2000) *Dollars, Distance and Online Education: The New Economics of College Teaching and Learning*. Phoenix, AZ: Oryx Press.
- Gladieux, Lawrence, and Watson Scott Swail. (1999) *The Virtual University and Educational Opportunity: Issues of Equity and Access for the Next Generation. Policy Perspectives*. Washington, DC: The College Board.
- Hanson, Dan, Nancy J. Maushak, Charles A. Schlosser, Mary L. Anderson, Christine Sorensen, and Michael Simonson. (1997) *Distance Education: Review of the Literature, 2nd Ed*. Washington, DC, and Ames, IA: Association for Educational Communications and Technology and

- Research Institute for Studies in Education.
- Howard, Caroline, Karen Schenk, and Richard Disenza. (2004) (Editors) *Distance Learning and University Effectiveness: Changing Education Paradigms for Online Learning*. Hershey, PA: Information Science Pub.
- Jenkins, Janet. (1989) "Some Trends in Distance Education in Africa: An Examination of the Past and Future Role of Distance Education as a Tool for National Development", *Distance Education*, 10, 1, 41-48.
- Lau, Linda K. (2000) *Distance Learning Technologies: Issues, Trends, and Opportunities*. Hershey, PA: Idea Group Pub.
- Lewis, Laurie, Kyle Snow, Elizabeth Farris and Douglas Levin. (1999) (Editors), *Distance Education at Postsecondary Institutions: 1997-98*. Washington, DC: U.S. Department of Education. National Center for Education Statistics.
- Mills, Roger. (2003) "The Centrality of Learner Support in Open and Distance Learning: A Paradigm Shift in Thinking", in Alan Tait and Roger Mills (Editors), *Rethinking Learner Support in Distance Education: Change and Continuity in an International Context*. London: RoutledgeFalmer.
- Olsen, Alan. (2003) "E-learning in Asia: Supply and Demand", *International Higher Education*, 30, Winter.
- Perraton, Hilary. (2000) *Open and Distance Learning in the Developing World*. London: Routledge.
- Peters, Otto. (2001) *Learning and Teaching in Distance Education: Analyses and Interpretations from an International Perspective*. London: Kogan Page.
- Petrides, Lisa A. (2002) "The Internet in Higher Education", in James J.F. Forest and Kevin Kinser (Editors), *Higher Education in the United States: An Encyclopedia*. 2 volumes. New York: ABC-CLIO Publishers, 367-373.
- Phipps, Ronald and Jamie Merisotis. (1999) *What's the Difference? A Review of Contemporary Research on the Effectiveness of Distance Learning in Higher Education*. Washington, DC: The Institute for Higher Education Policy.
- Rumble, Greville and Colin Latchem. (2004) "Organizational Models for Distance and Open Learning", in Hilary Perraton and Helen Lentell (Editors), *Policy for Open and Distance Learning*. London: RoutledgeFalmer.
- Snydman, Stuart K. (2002) "Technology", in James J.F. Forest and Kevin Kinser (Editors), *Higher Education in the United States: An Encyclopedia*. 2 volumes. New York: ABC-CLIO Publishers, 652-659.
- Tait, Alan. (1997) (Editor) *Perspectives On Distance Education: Quality Assurance in Higher Education*. Vancouver, Canada: Commonwealth of Learning
- UNESCO. (2002) *Trends, Policy and Strategy Considerations*. Paris: UNESCO.
- Werry, Chris. (2001) "The Work of Education in the Age of E-College", *First Monday*, 6, 5, May.
- Williams, Marcia L., Kenneth Paprock, and Barbara Covington. (1999) *Distance Learning: The Essential Guide*. Thousand Oaks, CA: Sage Publications.

Websites

- Campus Computing Project.
www.campuscomputing.net
- Commonwealth of Learning. www.col.org
- Distance Education at a Glance.
www.educause.edu
- International Center for Distance Learning
www-icdl.open.ac.uk
- Western Cooperative for Educational Telecommunications. www.wiche.edu/telecom

Institute for Higher Education Policy.
www.ihep.com

James J.F. Forest
Combating Terrorism Centre
United States Military Academy
West Point, New York, USA
James.Forest@usma.edu

Education Policy: Preschool

Edit Andrek

Introduction

Historically, the predecessors of pre-school establishments were orphanages, since they were originally established to fill a need in the care of parent-less children. With the industrial revolution in the 19th century and the growing requirement for an additional labour force, which led to the employment of women, who until then traditionally took care of their children full-time, the early pre-school institutions' remit included care for children living with their working parents. Further urbanisation and migration led to the disintegration of large, extended families where senior members of the family (grandparents) played a crucial role in bringing up the younger generation. Pre-school institutions were, for a long time, perceived as an establishment for "taking care" of children temporarily. Consequently, their work was initially guided more by medical, than pedagogical considerations.

A major change came about in the mid-19th century when, influenced by the advanced knowledge of a child's emotional and physical development, the focus shifted first from children's physical health and establishing habitual routines, to more comprehensive socio-emotional development. The focus later shifted to cognitive development and development of the capabilities necessary for securing success in school (educational process). Past developments demonstrate that the relationship towards children and stressing certain educational values have been historically initiated, as a result of complex social, pedagogical, psychological and political changes that led to the changes in societal approach to child care. These changes in society led to changes at the micro

level in the family. The state has resumed a part of the responsibility for child care, which was necessary to ensure that the younger generations are more socially responsible and can fit in well within society as a whole.

Pre-school education lays the foundations for personality development and the quality of pre-school education largely influences the quality of further education. However, the social character of education is twofold: the family is the basic social cell that has a leading role in bringing up young generations, whilst societal survival depends on the quality of their education (see Kamenov 1999). It is now the right of every child to have an institutional pre-school education and society is obliged to secure it. When the state organises pre-school educational institutions, it has to take into account children's needs, such as the need to initiate intellectual development, to broaden social contacts, to play and to take part in physical activities etc. where the institutionalised education will be complementary to 'family education'. In order to realise the planned activities, there must be close collaboration and mutual understanding between parents and pre-school teachers. This interface requires a good knowledge of children's socio-cultural backgrounds and taking these into account when preparing the curriculum. There is no uniform pre-school education model, but the curriculum has to be 'locally coloured', taking into account the specifics of the local/regional setting from which the children come (see Woodhead 1979).

Nevertheless, social education is not static. As society is in the process of constant change, so then must pre-school curricula be flexible enough to adjust to the social needs. In order to have good quality curricula, it is necessary to secure the collaboration of all stakeholders in the process of curricula development at the different levels of policy

formulation and decision-making. In fact, the most successful have been the compensatory early education programmes that included work with the family and where the family had full social support to take over the responsibility for the development and advancement of its children.

Historical and Social Factors

Ideas dealing with bringing up and educating children can be found in the works of the ancient Greek philosophers. Plato talked about the notion of education and its importance for creating an ideal state (republic). The concept of the importance of education advancement has remained until today and is very dependent on political, social and economic factors and national and social traditions etc. These factors directly and indirectly influence the creation, regulation, curricula, pre-school teachers' training and mode of financing of the pre-school educational institutions and systems. It has not been contested that the first few years are of crucial importance to the development and quality of human personality and society recognising this is, in fact, the first necessary step in deciding the format and content of pre-school education programmes.

A number of well-respected scholars dealt with the issue of bringing up children, from John Lock, Rousseau, and Pestalozzi, to the utopians and Robert Owen. The very first systematised pedagogical theory of early childhood education was offered by Froebel, and later Decroly and Maria Montessori. One of the major characteristics of early childhood is '*plasticity*' (that diminishes over the years). The process of the biological development of a small child is very intensive, where his/her nervous system and physical functions express a high level of plasticity i.e. their openness to influences of the environment, which are particularly strong and leave eternal marks on a child. The influence of the

social environment is particularly strong, especially if it is regular and systematic and can change a child's entire future development (Ilić in Kamenov 1999). The child is receptive to impressions coming from his/her immediate environment and in the first few years of life, he or she learns relatively more than in any other phase of his/her life.

The interest in early childhood is initiated by scientific research in brain development, as the grey mass of the brain is fully developed by the 6th or 7th year of life (Kamenov 1999). It is an assumption that if many useful associations between the neurons are not developed in very early childhood, they (the neurons) will atrophy. This is the point of departure in developing compensatory programmes of so-called "cultural deprivation", based on the assumption that unsatisfactory school performance and lack of motivation to learn are the consequence of modest experiences that children have attained in very early childhood development. Although children may have a chance to acquire quality knowledge later in life, the previously mentioned 'plasticity' at an early age is an important factor to be considered in building up good quality. It is perceived that pre-school education is important for developing young personalities and what the quality of learning in future life will be.

Bringing up children has a social character, not only because the family is the basic social cell, but because a good quality of education will influence the overall quality of life and overall will influence the development, and even the survival, of society. This may be the reason for supporting the family in discharging its functions in bringing up children. But, then again, historically, kindergartens opened their doors to accommodate day care for children of mothers who joined the labour force at the

time of the industrial revolution in the early to mid-19th century.

Functions of Pre-School Education

The functions of pre-school education are as follows:

1. To ensure meeting children's and societal needs and the realisation of their rights. These state that children have the right to be born and live in an environment that will enhance their physical and mental health, where they will feel accepted and loved, in which they will have the best conditions for growing-up and for the development and learning that society can secure for them, with no exclusion, segregation and/or discrimination;
2. To secure active inclusion in the life of the children's communities;
3. To prepare children for school education;
4. To release parents from caring for children for part of the day;
5. To assist the family in bringing-up (social education) children and enhancing the level of pedagogical and psychological culture of parents;
6. To discharge compensatory functions of the out-of-the-family pre-school education;
7. To assist and protect children with special needs and children from under-privileged social backgrounds, and
8. To assist gifted children in realising their real potential (see: Kamenov, 1999).

In assessing the programme and possibilities of influencing the development of children in a pre-school institution, one must bear in mind the depth of family influence on a child's development. The experience of "compensatory programmes" has shown that the most successful were those that were based on inclusion in the process. The effects show to what extent parents are included in the realisation. Each and every programme that was marked

"successful" reported good collaboration between teachers and parents. On the other hand, the strength on intra-family relations and the intensity of family influence on the child shows that pre-school education can only be supplementary to good family education and not a replacement. The socio-cultural context in which the child is growing-up would be the point of departure in defining the curriculum.

At the beginning of the 1960s, in the US, the idea emerged of changing society through education, where education would assist in reducing poverty, decreasing unemployment and helping with other social shortcomings. The "compensatory programmes" emerged to assist children from under-privileged backgrounds through early 'societal intervention'. These programmes assume that poor social conditions are a predominant factor for children's underperformance, as the children have been exposed to negative experiences by the family and the environment. Pre-school education institutions, with their organisation, rich programmes and quality of activities will partly diminish these.

Unfavourable socio-cultural and family conditions can be weakened through the stay of children in an institution where an appropriate "compensatory programme" is on offer. Although these attempts were interesting and more or less successful in different settings, practice has reiterated the strong influence of the family on a child's development and that the comprehensive pre-school education programmes have to incorporate actions that will influence the family and family behaviour. The compensatory programmes may not have delivered results in the long run, but they certainly had some noticeable short-term results, such as higher employment, lower level of truancy, fewer pupils who discontinued their studies, lower crime rates

and increased home ownership etc. The investment in the education of parents and the improvement of children's living conditions appeared to be the most important for the overall success of the "compensatory programmes".

It has also distinguished between the problems that emerged due to cultural differences and those that were initiated due to bad societal conditions in which a child was growing up. Shortcomings are a discrepancy between what one expects from a child and what his or her real abilities are and that is why, in the process of curriculum development, the developers should begin with what a child knows and not with what he or she does not know. More emphasis should be placed on preventive activities and not on the eradication of real and supposed errors. Individual work with children, taking care of a child's individual characteristics and developing a curriculum, based on detailed knowledge of targeted children, (group) is probably the best way to enrich a child's personality and experience.

Liaison between pre-school education institutions, parents, primary schools and the community ensures that the child is exposed to many factors, emphasising their continuing experience, not only in learning, but in the overall approach to forming their personality. Continuity in educational and intellectual experience plays an important role in facilitating children's confidence in their immediate environment. Matching expectations and the consistency of the reaction to children's behaviour, gives children the self-confidence they need to interact socially within both their social and intellectual environments. Continuity, in a strictly educational sense, enables children to have a continuous learning experience and multi-aspect development, engaging both the immediate environment and wider social community.

Bronfenbrenner has developed a model of "ecological intervention", which assumes the major changes in a child's environment and the reaction of the person in charge of that child's development. The Council of Europe has taken the position that pre-school education must be directed towards the welfare of a child. The primary role of pre-school educational institutions should be education and socialisation, and not purely the 'housing' of children. It should also be emphasised that pre-school education is not extended to the primary school curriculum, but has a number of intrinsically specific principles, contents and methods of dealing with children of pre-school age.

A number of authors have studied the importance of the curriculum, from various perspectives. "The curriculum study showed us that diverse curriculum models can be *equally* effective in improving children's education and that this success does not appear to derive from the curriculum models themselves, but rather from the way programs are administered and operated" (Schweinhart 1986:40). Stephens (in Hoffman 2003:204) stated that "childhood today, as situated in a social and cultural construct, is also a political one: cultural ideas can be and are used to frame policy, and are legitimised by educational and training institutions, policy makers and others in position of power. While we cannot assume isomorphism between discourse and practice, at the same time, discourse exerts a powerful role on defining normative and "best" practices with regard to policy and practice".

The societal realisation that work with pre-school children is important for society as a whole, initiated the need for the introduction of educational content in work with pre-school children. The period between the third and fifth years of life is important for the beginning of learning (Boocock 1995). Although it is logical that early childhood

provisions differ, depending on the state of a country's economic development, the fact is that between countries of the same level of economic development, there are differences in the level of development of pre-school educational institutions. Comparative analysis shows that there are different levels of government engagement in pre-school education, ranging from full to marginal financing and from full to loose control over curriculum development and implementation etc. As stated, "Only recently has systematic international research begun to document the linkage between national policy, early childhood programs, and outcomes for children" (Boocock 1995:95). Comparative research has also demonstrated that there are large discrepancies in the relationship between empirical research and various government policies.

Boocock (1995) singles out France as the country where research influences policy decisions significantly. In France, a large scale research was conducted between 1983 and 1990 on the effect pre-school experience may have on later success in schooling. As a result of these findings, nurseries in France are attended by all children between three and five years of age. There are an increasing number of children younger than three, who are taking part in pre-school education. The whole-day programme is partly academically oriented and focus has been assured by giving pre-school teachers the same status, same quality training and salaries as primary school teachers.

All education employees have civil service status. Research has proved that every year spent in a primary education institution reduces the possibility of failure in school, especially for children coming from underprivileged social backgrounds (see Boocock, 1995). Comparing the French results to similar work conducted in the UK and Germany, the author concluded that in all

countries, no factor was found that proved more influential on a child's educational performance than attendance of an organised pre-school programme. There is a consensus amongst professionals that Sweden has the most effective system of child care in the developed countries. As there are a large number of employed mothers, the system of pre-school education institutions is highly developed.

A survey in Sweden showed that the participation in a pre-school education programme before the first year of a child's life led, not only to their improved verbal abilities, but also improved their tenacity, independence, anxiety levels and self-confidence, compared to children who joined the education process later (Boocock 1995). This research project has, surprisingly, shown that putting children into educational institutions at an early age not only has no negative consequences, but in fact, improves the child's performance. However, it should be remembered that in Sweden, the level of uniformity in the education system is fairly high, as the government, through various regulatory initiatives, centralised financing and controls, and they ensure that the quality of work is within the benchmarks imposed by the government, which is an ultimate guarantor of good quality education before the electorate.

Japan is another developed country that experienced unprecedented growth rates and was successfully transformed from a feudal society into a post-modern society that can boast many societal achievements, over the last hundred and fifty years (or so). It is a fact that education has contributed to that rapid development (in a unique blend of traditional Confucian values and philosophy of life) and that it proved to be the major factor for socialisation of children. The *Amae* relationship between mothers and children, which assumes high protection and over-

indulgence, is significantly different from the rigidity that dominates society. Kindergartens are the first and by far the largest, most important step in the process of socialisation of a child and to his/her belonging to a wider group. This is why institutions are important in Japan and why children are closely involved in the work of those institutions. Policy makers are constantly re-examining the values promoted and the quality of pre-school educational programmes.

A curriculum that has a purely academic content is not dominant at all, at least not based on the positive instructions and published regulations enacted by the Ministry of Education. The focus is on developing group spirit and societal co-operation, ensuring that conflicts between peers are solved amongst themselves, whilst the teacher remains the neutral, non-threatening figure, who is there to organise a simulating environment and nothing else. The main aim is to develop a positive attitude towards school amongst children and if this is done, the aims of the curriculum are easier to realise. In Japan, the content of the pre-school curriculum is divided into five main areas: 1) language, 2) environment, 3) interpersonal relationships, 4) expression (ability to express oneself) and 5) health (physical and emotional well-being).

Preschools in Comparative Perspective

Boocock (1995) mentions that in Singapore, there are kindergartens that follow the programme developed or endorsed by the ruling party, although there are also kindergartens following the programmes offered by both the opposition and independent parties, with specific privately developed curricula. Singapore is another Asian country that experienced very rapid growth. It is a country where families are decreasing in size, with the average family having one or two children. This is why,

similar to Japan, the emphasis is on group experience (experience living and working in a group) and sharing and co-operation; something that children cannot easily experience in an atomised family (lacking the experience of living in a large (extended) family).

In Australia, prior to the 1970s, there was a project conducted between people who were allocated accommodation under the auspices of the low income housing project. The results of this research have shown that children who were included in pre-school education programmes demonstrated some advantages, which surprisingly disappeared by the end of year one of primary school. However, the project has indicated another positive consequence i.e. mothers who participated, organised a number of extramural social programmes and welfare activities of their own accord (Boocock, 1995). The results were important, as they also disclosed that the mothers' satisfaction with life (style) rather than job satisfaction played an important role in the socio-emotional development of the child.

Research conducted in New Zealand has shown that the inclusion of children in educational programmes had the following consequences on families: an improved relationship with children, a decrease in the level of mothers' stress, an upgrading of education or training credentials and improved employment status (see Boocock, 1995). These findings have established a need for pre-school education and offered policy makers sufficient information and suggested ways in which the current puzzles can be solved. All the results and survey findings have demonstrated the need for strengthening institutional forms of pre-school education in various countries.

In contrast to developed countries, where the efforts of policy makers have been oriented towards improving the quality of

pre-school programmes since the number of children involved is high, in under-developed countries the major problem is malnutrition and a poor health record of students. Therefore, in underdeveloped countries, the primary task of pre-school educational institutions is to help solve these problems, as well as stimulating the cognitive, psychological and social development of the children. In Columbia (see Boocock, 1995), recent research has shown that children who participated in an early start to pre-school education, achieved higher levels of cognitive development and this was directly correlated with the length of time spent in pre-school educational institutions. The increase in IQ was constantly supported until the eighth year of life.

It has been reported that in India, around only 50 per cent of children attending primary school complete the first grade, which is worrying. This is connected to the growing demand for labour and the general trend of employing children as cheap labour. Girls are often left at home to take care of other children, especially their brothers. This is why there is very little engagement of and with, pre-school educational institutions. The usual problems of malnutrition and poor health in children have also been found in India, especially in the northern states. In order to address these problems, the Government of India launched the Child Development Service in 1975, where professionals spent a few hours every day with children, working on developing their skills, improved health control and providing a steady daily meal (Boocock 1995).

The results have shown that it is possible to reduce the mortality of children, improve malnutrition and spur academic advancements in future schooling. Therefore, the advice to policy makers is to consider a combined approach which encompasses different aspects of nutrition, health and

educational programmes, and measures the achievement of lasting successes, not only in the physical but also in the intellectual sense. The combined approach is the best for under-developed countries. It is also useful to become familiar with programmes in other, primarily developed countries, ensuring that good practices are considered for application in developing countries. This is where comparative education and comparative education policy steps in. Although the state of development of the respective countries may differ, the problems that are faced by policy makers may be similar, so the experiences of other countries may prove to be very valuable in developing innovative solutions. Research has demonstrated that the inclusion of children in good quality pre-school education programmes has some advantages, which are reflected in the achievement of better academic results.

The most advantageous children are those in countries that have a very well-developed national policy, which secures the pre-school educational experience for the entire child population, with a long experience of having organised and managed pre-school educational institutions, and where the law regulates most if not (almost entirely) all aspects of the pre-school educational process (including the normatives of the building, equipment, fixtures, curriculum, learning objectives and outcomes, training and education of pre-school teachers, supervision and control of pre-school educational institutions, evaluation the process and outcomes, etc.).

Policy decisions are quite often politically coloured, despite the results of much of the empirical research. Although education is “the future of a country”, the degree of social investment and consistency in political and economic sense of government policy, limits the degree of inclusion and the number of children taken through the pre-school

educational programmes. It may be surprising to some, but politics is by far the most important factor in the educational decision-making process. Early childhood services (ECS) in New Zealand have a long history, which stretches over 120 years and includes both private and community-owned institutions. The government has subsidised some of these institutions since the beginning of the 20th century and private kindergartens have been supported from budgetary sources since the last few decades of the last century. The government only gives grant-in-aid to ECS and thus administers policy for ECCE, but does not administer the services themselves: committees, boards or owners carry out the day-to-day management and administrative functions (Meade 2002:6).

The movement for Maori families (*Te Kohanga Reo*) was launched in 1982, with the main aim of strengthening Maori families and protecting the Maori language. From the beginning of the 1980s every policy decision has had to take into account the consequences for Maori children, as descendents of the original settlers of the land. The national curriculum for children from birth to the fifth year of life was enacted in 1996. The title of the curriculum is *Te Whaariki*, which in Maori language means “A woven mat for all to stand on” (see Meade 2002). The curriculum has four comprehensive principles: empowerment, holistic development, family and community, and relationships. What facilitates the achievements of the proclaimed results are: belonging, well-being, exploration, communication and contribution.

Societal Importance

The importance of early education must be observed when one considers the training of pre-school teachers and determining their status. In France and Norway, the status of pre-school (early childhood) teachers is equal

to the status of primary school teachers, with their education and training requirements being the same. Taking into consideration the complexity of the relationships in early childhood and children’s needs, it is necessary to secure good quality and highly trained personnel, capable of meeting the needs of modern times; but who will be equally well remunerated. The activities of pre-school teachers cannot be limited to contact with children, but an equally good rapport has to be established with the parents, the local community and other educational institutions etc. This requires a more complex education and training curricula for pre-school teachers and longer schooling, facilitating the specialisation and development of in-depth abilities in designated areas of training.

In Sweden and Norway, families with young children receive financial support. In fact, in Sweden every child whose mother is employed or is in full-time education will have a secured place in a kindergarten or other pre-school educational institution. Children in Sweden begin their schooling at seven years of age, but almost all children attend the “pre-school classes” which follow the school curriculum (Alvestad and Samuelson 1999). About 60 per cent of all those employed in Swedish kindergartens are university educated. Since a high percentage of mothers are employed, society is aware of the importance of early childhood for future development and has taken care of children from a very early age. Pre-school experience is secured, not only for children whose parents are at work, but also for children of unemployed parents.

This change came about in 2002 with a change in the pre-school regulations (Taguchi and Munkammar 2003). In theory, the importance of early childhood experience for later life-long learning has been strongly emphasised in recent times. Kindergartens are

perceived as being separate from school and the learning process, but are seen as a point of departure in the process of life-long learning. The legislative changes promulgated in 1998 ensured that in Sweden, pre-school education is now part of their unique education system. In the same year, the first national pre-school curriculum was enacted, which formally equalised development and learning. The document underlines not only the social importance of proper child development, but also proper learning. Policy makers did not want to follow or mimic the primary school programme and activities, but rather, they wanted to ensure that the first year in the primary school is, in fact, influenced by the new pre-school curriculum.

Pre-school educational institutions are perceived to be the starting point in the process of long-life education and this is why the focus gradually shifted from welfare to learning, when curricula are defined. The aim is to facilitate transition between the two institutions. This is obvious in and for Sweden, where the structure of pre-school curriculum for children up to five years of age is very similar to the curriculum for children between five and sixteen. The curriculum still includes goals, aims and objectives, perspectives on learning and values etc., but the working methods are not set out and evaluation is carried out through a full documentation of activities (portfolio building).

The deregulation process, characteristic to Sweden, is also applied to the education process and decisions are made in the classroom where both children and teaching staff are involved. Kindergartens are goal-oriented and it is left to every pre-school teacher to decide how these goals are achieved. In Norway, the curriculum is more detailed, with methods being listed, as well as the content, whilst the outcomes are clearly spelt out. Pre-school education, as part of the

overall education system in Sweden and in Norway, is regarded as part of the children's welfare system. The Swedish curriculum consists of norms and value, development and learning, children's influence, pre-school and home, and co-operating with the school, etc. (Alvestad and Samuelson 1999). This is explicitly listed as what children have to learn in a kindergarten.

The curriculum is often the result of political decisions and compromises. The goals have to be consistent with what kindergartens are capable of giving children, whilst on the other hand they have to be flexible enough to be accessible to each and every child. The Norwegian curriculum focuses on: society, religion and ethics, aesthetic objects, nature, technology and environment, language, literacy and communication, physical activity and health (Alveda 1999). It is very interesting to note that religion and ethics are first on the list, but this can be explained by the belief that Christian learning is the foundation of all societal ethical values. Strengthening the link with traditional values and reinforcing the role of the family in bringing up children are the most important features of this curriculum.

In contrast, the Swedish curriculum is dominated by the idea of democracy which is connected to the state, whilst church and religion are separate. This separation is clearly useful, as in multicultural societies it is impossible to pay attention to all religious learning. A holistic approach to the education of pre-school children claims that children learn by their senses. This approach develops equally all aspects of a child - physical, cognitive, emotional and social. As in the past, pre-school education was regarded as an important aspect of the welfare state, together with the idea that parents were not competent enough to play their parental role well. In 1996, pre-school education was transferred to

the Ministry of Education and Science and general social policy was transferred to educational policy (Taguchi and Munkammar 2003). Children are seen as a societal resource. They must be respected and enjoy their full rights (physical punishment is strictly prohibited). In fact, “Political policy-making has “married” issues of equality of women and children with society’s need for a well-educated workforce and lifelong learning, whilst building, maintaining and improving a welfare society for all citizens” (Taguchi and Munkammar 2003:30).

In Thailand, there is no separate pre-school policy, as the education policy begins with the school policy. As the primary interest of the state is public health, there is very little incentive to develop consistent pre-school education programmes. Mothers lack knowledge about child development and early learning patterns, as well as the importance of early child development on the later learning capacities. Children from three to five years of age are in the remit of the Ministry of Education, whilst children up to three years of age are in the joint care of the Ministries of Education, Public Health, University Affairs, Interior, as well as UNICEF and a number of non-governmental organisations (NGOs) (see Kamerman 2002).

Similarly, in Malaysia, a number of social players are involved in the process of pre-school education. The Ministries of Health, National Unity and Social Development, Rural Development and Education are all interested in children’s welfare. Early childhood education is conducted through home-based centres (usually for children under four years of age), and for children from four to six there are pre-schools. Pre-schools provide care for the children of working parents, stimulate cognitive, socio-emotional, physical development of children and prepare them for school (Kamerman 2002). The Ministry of Education prepares

the curriculum for all types of provision and is responsible for teacher training and the overall implementation of pre-school education. The Government, as a whole, is interested in early childhood education, following the introduction of the Act in 1996, under which all children of pre-school age are required to attend a pre-school. There is a general tendency to invest more in pre-school education and care. A number of policy decisions have been made to secure children’s welfare and indirectly their families, all through a good combination and coordination of the health and educational aspects of children’s needs.

In Indonesia, the development and children’s welfare is in the hands of parents and only if they are incapable of meeting their roles, does child care become an obligation for the extended family, neighbours and finally the care service. The latter has a limited role in supporting only the health and (physical) growth of children. There are also a number of institutions that are charged with supporting children’s development and children’s education; but these two functions have traditionally been kept strictly separated (Kamerman 2002). The “Family and Under Five Development Programme” educated mothers and other care providers on how to stimulate better children’s cognitive, socio-emotional and physical development (see Kamerman, 2002). The Government is in charge of teacher education and training, but the money allocated is very limited and cannot meet demand. The major impediments to the implementation of Early Childhood Care and Education (ECCE) policy measures are: poor health and difficult access to the health service, malnutrition of children up to five years of age, the overall poor quality of ECCE, as education does not focus on the needs of children and the position of women in society is marginalised etc. Governments also find it difficult to maintain the existing

number of pre-schools, which is an additional impediment to the successful implementation of pre-school policies.

In the Philippines, the mother plays the most important role in bringing-up children, and only when the extended family cannot take care of a child and a nanny cannot be afforded by the family, will the child (when over three years of age) be sent to a day care centre or private kindergarten. It is worrying that only 86 per cent of children will continue their education after the first grade of primary school, although primary education is *stricto lege* —compulsory. ECCE programmes are aimed at improving this situation, but immediate results are lacking. Public childcare centres and public primary schools are heavily subsidised by the government and the parents who cannot afford to pay the fees are required to work in the schools and centres. The costs of private kindergartens are often higher than those of university. The Early Child Care and Development Law of 2000 enabled the implementation of a new policy which focuses on children and parents, as the Law recognised the importance of early childhood education and re-emphasised the major role of parents as primary players in the process of bringing up and educating children. The law encourages public engagement with parents, through different educational programmes (Kamerman 2002).

In Vietnam, the Ministry of Education and Training, in collaboration with the Ministry of Health and Women's organisations, is charged with promoting programmes for children of up to six years of age. The quality of the offered programmes varies widely between urban and rural areas. There are also nurseries for children under three years of age and kindergartens for children between three and six. The curriculum is based on providing the cognitive, physical and social development of children. "Current policies are aimed at increasing supply and coverage

rate in kindergartens to 70-80 per cent, developing family day care homes for under 3's and stimulating public support and increased investment." (Kamerman 2002:20). Although government policies are aimed at involving an increased number of youngsters in the education process, the family is still regarded as the primary player in securing education and care for children, and there is an increased emphasis placed on their education.

In Cambodia, there is an emphasis on adequate daily nourishment of children, their health care and, finally, education. Children up to three years of age are either with their mothers who take them to the fields to work, or with their older siblings; while ECCE is focused on children from three years of age to school age. The Ministry of Education, Youth and Sports (MOEYS) formulated, for the first time, a policy on ECCE (Kamerman 2002). The Ministry of Women's and Veteran Affairs (MOWVA) is also charged with early childhood care and both ministries are involved in education and training of employed and volunteers, participating in various government funded or sponsored pre-school programmes. Approximately ten per cent of children attend pre-schools, while 25 per cent of six year olds are in kindergartens attached to primary schools. In Lao, there are two types of pre-school establishments: one for children under three and another for children from three to five years of age. The first group of institutions attracts barely 2 per cent of the targeted population, whilst the second attracts eight per cent (Kamerman 2002). The programme is focused on providing cognitive, physical and social stimulation of children and also on the preparation of children for further education in primary school. The Ministry of Education co-operates with NGOs to ensure the operations of both types of institutions, as explained above.

Conclusion

It is obvious that pre-school education programmes have to be responsive to ongoing changes in society. If parents are not aware of the importance of early childhood learning for the further development of a child, society must put appropriate policies in place to address this deficiency. It is necessary to use legislative tools to ensure that children's welfare is secured and that they have complete access to education, in order to become useful members of the community when they grow up. However, it is not only necessary to have the appropriate policies promulgated, but also to ensure that those policies are in place and – implemented.

Selected References

- Alvestad, M. and I.P. Samuelson. (1999) "A Comparison of the National Pre-school Curricula in Norway and Sweden", *Early Childhood Research and Practice*, 1, 2, 41-63.
- Boocock, S.S. (1995) "Early Childhood Programs in Other Nations: Goals and Outcomes", *The Future of Children*, 5, 3, 94-114.
- Calhoun, J.A. and R.C. Collins. (2001) "From one Decade to Another: A Positive view of Early Childhood Programs", *Theory Into Practice*, 20, 2, 135-140.
- Hoffaman, D.M. (2003) "Childhood Ideology in the United States: A Comparative Cultural View", *International Review of Education*, 49, 1-2, 191-211.
- Kamenov, E. (1999) *Predškolska pedagogija* [Pre-School Pedagogy.] 2 Volumes. Beograd: Zavod za izdavanje udžbenika.
- Kamerman, S.B. (2002) *Early Childhood Care and Education and other Family Policies and Programs in South Asia*, Paris: UNESCO, Early Childhood Policy Series 4.
- Meade, A. and V.N. Podmore. (2002) *Early Childhood Education Policy Co-ordination under the Auspices of the Department / Ministry of Education: A Case of New Zealand*. Paris: UNESCO, Early Childhood and Family Policy Series 1.
- Schweinhardt, L.J.; D.P. Weikart and M.B. Lerner. (1986) Consequences of Three Pre-school Curriculum Models through Age 15, *Early Childhood Research Quarterly*, 1, 15-45.
- Taguchi, H.L. and I. Munkammar. (2003) *Consolidating Governmental Early Childhood Education and Care Services Under the Ministry of Education and Science: A Swedish Case Study*. Paris: UNESCO, Early Childhood and Family Policy Series 6.
- Woodhead, M. (1979) *Pre-School Education in Western Europe: Issues, Policies, and Trends*. London: Longman, for the Council of Europe.

Edit Andrek
Independent Scholar
London, UK

Education Policy: Schools

Celina Su

Introduction

Schools are generally defined as places of learning, and the category of “schools” includes different institutions in different countries. For example, in some countries, “schools” only refer to institutions below the university level; in other countries, “schools” can refer to learning institutions at any level. This entry focuses on schools at the primary and secondary levels in the contemporary context of the late twentieth and early twenty-first centuries.

Over the past few decades the role of schools in preparing future workers for growing economies has taken precedence over their role in training future citizens. Growing debates surround whether and how local public participation and governance should take place, how school systems should be structured, and whether policies such as decentralization, school choice, standards-based reform help to render schools more efficient.

Large inequities remain in access to education by region, gender, race, and income within countries, and debates in the funding and structuring of school systems are often directly related to how governments address these inequities. Finally, bilingual education and emergency education are also educational policy issues of growing importance.

Schools as Incubators of Human Capital

The creation of human capital and knowledge is one of the primary goals of educational policy. An economic perspective of education focuses on the costs of public education and the subsequent returns to education yielded via increased productivity, most commonly measured via Gross Domestic Product (GDP). Some estimate that “investments in human

capital over the past two decades may have accounted for about a half a percentage point in the annual growth rates of those countries”, and slow economic growth in countries such as Egypt and Tunisia is at least partly attributed to their low educational attainment (UNESCO 2005).

According to this logic, investing in higher quality schools yields better-trained students, who are, therefore, more productive in the labor force and, on a macro scale, jointly render the entire country’s economy more productive (Mankiw 1995). Most international institutions also emphasize economic productivity as the main goal of education. For instance, the bulk of the World Bank’s current education efforts are labeled “Education for the Knowledge Economy” and focus on programs to “equip countries with the highly skilled and flexible human capital needed to compete effectively in dynamic global markets” via improved tertiary education and greater access to science and technology.

A look at the economic literature from the 1970s through the 1990s provides one possible reason for the rise to prominence for human capital theory. In the 1970s, the wages of university-trained white males, for example, were not much higher than high school graduate white males. Returns to education were low, and some even lamented the “overeducated American” (Freeman 1976). By the 1990s, the rise in prominence of information technologies and computers appeared to produce higher returns to education. The manufacturing sector, which in industrialized countries is associated with high unionization rates and higher wages than those in the service sector, concurrently shrank. Thus, the wage gap between college and high school graduates increased, and overall wage inequality increased (Levy and Murnane 1992). In Paraguay, a person with tertiary education earns as much as 300% as

one with secondary qualifications (UNESCO 2005).

In developing countries, governments tend to focus on building high rates of numeracy and literacy. These skills enable workers to engage in entrepreneurship and obtain more complex and better-paying jobs. This has led to *brain drain*, or situations in which highly educated persons emigrate to other countries. To offset these effects, developing countries aim to increase educational attainment of significant portions of their population and grow domestic economies. Other strategies include failing to fully license the professionals they train, so that they are not eligible for licensed work in other countries. The overall impact of brain drain is mitigated by the overall benefits of education. Education measures are strongly correlated with health outcomes, lower crime rates, and less dependence on social welfare systems. Therefore, it may build capital in various ways.

Tracking in Education

Because of the strong correlation between education and income, governments attempt to match what is being taught in public schools with forecasted needs in a growing economy. Yet this dominant perspective of schooling has been criticized for reproducing hierarchical class structures. Economists such as Bowles and Gintis (1976) and sociologists such as Bourdieu (1979) suggest that the public education is structured as to mold working-class students into future workers. They argue that in public schools students perform drills, give authority to teachers standing in the front of classrooms, and shift tasks at the sound of a bell; this is contrasted with primarily private, elite schools where students are encouraged to ask questions via a Socratic method, learn critical thinking skills, and engage in creative assignments. While Bowles and Gintis later critiqued their own

work as to leave greater room for student agency (1980), their 1976 inquiry into the actual social outcomes of education systems remains influential. Bourdieu (1986) explicitly describes *social and cultural capital* transferred by teachers in schools; in addition to skills embodied in what is called human capital, schools also transfer different levels and types of social capital (norms and resources of membership in social networks) and cultural capital (knowledge and attitudes) that enable students to better navigate privileged environments.

In Latin America, Paulo Freire (1970) led a popular education movement that emphasized *conscientização*, or consciousness-raising, in adult literacy and secondary school education, by teaching students critical thinking skills as well as basic reading and writing. Based upon his experiences teaching in Brazil and Chile, Freire's teachings and policies emphasized the active, participatory roles of students and adults in shaping their own education.

Within public schools systems, it can be argued that class-based tracking is also part of policy. In most European countries, for example, children must choose and take examinations for the secondary schools they wish to attend. Many of these secondary schools are vocational (*education policy: vocational and beyond*) and provide *apprenticeships* in blue-collar or trade careers, while others are more academically oriented, attempt to prepare students for university entrance, and are intended as training for white-collar or professional careers. Policy questions remain, however, regarding whether there are universal curricula all students should learn, and what these curricula should look like.

In developing countries where rates of secondary school attendance are lower, such tracking is less likely to be a pressing issue. Regional governments, especially those in poor areas, might also resist demand for

education if they feel that better educational services will repel businesses that rely on low-wage labor. Tandler (2002) explores such a case in northeastern Brazil, where businesses stated that they feared workers would become ‘uppity’ or recalcitrant if they became more educated, so they preferred firm-specific training to transferable skills. Governmental officials, in turn, felt that their competition with richer regions trapped them in a vicious cycle, whereby they risked losing geographically mobile workers and business investment if they invested in better provisions of educational services. Supposedly compensatory policies, such as subsidizing large firms’ private education programs, “distract attention from the problem of improving basic education” (ibid., p. 43). Such problems are perhaps inherent to labor-focused policy. However, it is often forgotten that education confers a host of other benefits.

Schools as Centers of Citizenship

Class-based criticisms of public education, such as those by Bourdieu and Freire described above, make explicit the role of schools as more than incubators of human capital. Specifically, schools are also important centers of citizenship training. It is in public schools that students learn the norms, civics, political systems, and language official to the country in which they live. In many schools, the day begins with students chanting a salutation to the national flag or singing the national anthem. Some education policy debates, then, center on issues of *multiculturalism* and the characterization and representation of different countries and minority populations in the lessons, textbooks, and literature used in schools. Policies regarding free speech, academic liberty, and dissent are also related to the degree to which public schools, especially those at the secondary and tertiary levels,

promote notions of citizenship as defined by the government.

At its extreme, public citizenship training in schools is seen as indoctrination. Such indoctrination is especially controversial in societies with *undemocratic governments*, such as the German Third Reich and Cambodia during the reign of the Khmer Rouge, where it may be easier for governments to distribute propaganda through public schools.

School Governance

School governance concerns the decision-making bodies and processes that dictate how funds for a public school are spent, what personnel are hired, what curricula are taught, how students are admitted, and how school activities are structured. In most school systems within industrialized nations, there are official bodies of school governance, often called School Boards, Parent Teacher Associations, or Local School Councils. These vary greatly in the amount and scope of real decision-making power they hold.

In highly centralized school systems, most policy decisions are made by the country’s central government. In these systems, parents, community groups, nongovernmental organizations, and other *stakeholders* attempt to influence school governance by appealing directly to national governments for more funding, the implementation of specific policies, or legislative changes. For example, both proponent and opponent groups lobbied members of French Parliament before they passed a law banning religious symbols, including veils worn by female Muslim students, from public schools. In this context, local and school-level bodies such as Parent Teacher Associations tend to have little power within the schools their children attend. Rather than helping to hire and fire school principals and teachers, for example,

these bodies aid schools in structured activities such as fundraisers.

Proponents of *participatory governance* argue that parental and community involvement is highly correlated with higher levels of education attainment, empowerment in community development, and greater legitimacy in the decisions made. They, therefore, tend to favor decentralized systems. A prominent example of institutionalized participatory school governance exists in Chicago, where elected parents and neighborhood residents, teachers, and principals participate in Local School Councils and oversee budget allocations and administration employment (Fung 2004).

Some policymakers warn that such decentralized structures are vulnerable to corruption, and that without centralized support, such decentralization also exacerbates inequities between low- and high-income school districts. In Argentina, a 1993 law created local School Site Councils in what had been a highly centralized educational system (Pini and Cigliutti 1999). The local and provincial governments were forced to assume fiscal responsibilities that were previously supported by the national administration. Without guidelines or oversight, some teachers' salaries and educational investments decreased, and administrators protected their interests by preventing parents from participating meaningfully in School Site Councils. The more successful parents were those who also brought money to the schools and came from high-income households. Critics state that parents and teachers were frustrated over the apparent disconnection between the Ministry of Education's political discourse about the "democratization of educational institutions" and actual outcomes in school quality, work conditions and teacher's salaries (*ibid.*, p. 199).

Similar outcomes were seen in Chile when the system was decentralized in 1980. There, responsibility for continuous quality improvement in teacher training was transferred from professionals within the school system to local officials who had no formal training, resulting in a decline in the quality of education, especially in smaller localities.

Fung and Wright state that outcomes are best when decentralized power is accompanied by higher-level support to formal participatory institutions, and parent involvement is most effective when concrete problems and solutions are publicly deliberated (2001). In clientelistic scenarios, for instance, increased resources can still help children learn when they are in the form of directly distributed school supplies (Easterly 2001); governments can install oversight boards such as those in the Chicago case, and non-board member parents can serve as divested monitors.

Governance and the Labor Market

While education is almost always a social policy priority, Pritchett describes other larger political economic contexts where low reported economic returns raise funding concerns (1999). Specifically, these scenarios involve (a) schooling of such low quality that it fails to bestow students with human capital, (b) economies where the supply for skilled labor exceeds demand, and (c) education attainment that is actually associated with negative economic growth rates. While these scenarios appear to be especially gloomy at first glance, they are fortunately associated with specific sets of reform policies. The first situation dictates the greatest role for parent-driven reforms, for they can work with teachers and lobby for resources and curricula which teach students work-related skills. The second and third situations call for supra-local policies. For instance, Pritchett writes that the

second scenario, where skilled labor exceeds demand, help to explain low returns to education in closed African regimes (ibid.). Nonetheless, parents and educators still have a role in school reform because they can help to develop adaptive strategies that allow educated workers to adjust to sectoral shifts in the economy, as well as encourage geographical mobility among young workers.

While school governance structures and experiences vary greatly both within and across countries, some thematic lessons emerge across case histories and settings. Centralized systems can become bureaucratic and unresponsive to local needs, but decentralized systems can have less knowledge infrastructure, and can suffer from clientelism. Parents must have the capacity to articulate demands and help implement change, and this capacity depends on sufficient funds and training, regardless of the parents' socioeconomic backgrounds.

Access to Education

In industrialized countries, the existence of universal, free primary and secondary education renders inequities in access to education less straightforward than those in developing countries. Most industrialized countries boast of high enrollment rates and have laws banning the involuntary segregation of students by race or gender in public schools, though some local laws banning illegal immigrants from public schools constitute notable exceptions (Armbruster et al. 1995). Therefore, much of the critical discussion on unequal access to education lies under the subheadings of "Tracking in Education", "Debates in Funding Education", and "Structuring Education." The two main types of inequities in access to education concern 1. differentials in funding levels, and 2. differentials in the type of schooling provided to students, with

relatively little access to elite, pre-professional, and pre-university education.

Regional and local disparities in access to education are especially pronounced in countries where school funding is decentralized and national-level funding is low. In the United States, for example, most public school funds come from local property taxes. Funding disparities are exacerbated in a system where sub-populations marked by race, ethnicity, and income often live in segregated communities. As a result, poor, primarily African-American and Latino neighborhood schools do not receive enough federal funds to fully compensate for the low levels of local funds generated by local property taxes.

For the most part, educational policy in low-income countries is focused on basic enrollment and addressing inequities in access, while focus in middle- and high-income countries centers on issues in governance and administration. In developing countries, basic public health infrastructure and education are also discussed jointly, while health policy is more likely to be debated separately in the context of industrialized countries.

The United Nations Educational Scientific and Cultural Organization (UNESCO), the World Bank, and international non-governmental organizations are prominent in helping developing countries to set educational policy. Official United Nations Millennium Development Goals, ratified by 152 countries in 2000, include universal primary education for both boys and girls by 2015. For the most part, emphasis lies in school construction and the hiring of teachers in order to increase overall access to education.

Large inequities in enrollment exist along gender lines, particularly in secondary school. While these disparities are commonly attributed to gender discrimination in

families, who may value educating sons more than educating daughters, there are in reality multiple contributing factors. Because families often fear for their daughters' safety, girls' enrollment often surges dramatically when a school is constructed near enough so that they can attend school and return home every day. Sometimes, simple provisions such as gender-segregated outhouses can also increase girls' enrollment. In other contexts, however, girls do not receive enough protection from sexual harassment by teachers or fellow students in the schools themselves. In poor regions, health and nutrition policies can also dramatically increase the effectiveness of educational policies aimed at increased enrollment. For example, Bolivia and Bangladesh, listed by UNESCO as reporting high rates of girls' school enrollment, made substantial progress from 1995 to 2005 through policies such as culturally appropriate campaigns to encourage girls' participation in rural indigenous communities, eliminating school fees for girls at certain levels, and providing food in exchange for continued enrollment.

Some researchers have argued that since universal access to education is simply impossible in the short-term in low-income countries such as Malawi, policymakers must focus on quality and other immediate goals as they continue efforts to increase access and address inequities in the long-term. For example, they must ensure that students who do enroll in school stay long enough to achieve minimum literacy, that implementation schemes are adjusted to fit different situations, and that parents are encouraged to participate and help to speed up policy change (Chimombo 2001). One danger lies in providing a comprehensive education for a very small percentage of the population; some developing nations have arguably devoted funds to secondary and post-secondary education, for example, at the

expense of primary education. In 1998-2000, whereas 68.6% of Guatemala's education spending went to pre-primary and primary education, only 7.2% of Armenia's and 22% of Mongolia's did so (UNDP, Human Development Report 2003).

Another manifestation of the tension between perceived quality and equity in developing countries is that of technology use. While it can be argued that the quality versus equity equation is not a zero sum game, high-cost technologies such as personal computers rarely increase educational attainment without professional training. In the meantime, lower-cost technologies such as radios are already familiar to most people; such investments can facilitate both local instruction and distance learning without professional training, especially in rural areas with little technological infrastructure.

Debates in Funding Education

The adequate funding of public schools may be the most prescient challenge in any country's education policy. Education funding is also the main factor contributing to persistent inequities in access to schools. Because most returns to education are long-term, it requires political will to fund education. This is especially true in countries with decentralized or federalized school systems, where the percentage of public funds coming from the national government is relatively low. In such contexts, debates in funding education are mostly conducted at the state and local levels.

In developing countries, the need for increased education funding is generally clear and uncontroversial. Some policymakers argue for debt relief by institutions such as the International Monetary Fund (IMF), arguing that many low-income countries spend three to four times on paying off debt what they spend on health and education services. National governments, rather than

local ones, usually pursue programs toward universal education.

In industrialized countries, education funding policy is more controversial because rudimentary school systems have already been developed, and policymakers disagree as to the continued quality of public schooling necessary, as well as how public funds can be maximized. Advocates of neoliberal economic policies argue that more funding does not increase returns to education; rather, more funding feeds ever-increasing bureaucratic structures surrounding educational services. In contrast, those who advocate more government funding in public schools argue that better pay for teachers, more classroom equipment, and smaller class sizes have been shown to improve outcomes in experimental settings.

Economic studies are more contentious in wealthy countries more likely to focus on neoliberal policies, such as the United States, where a smaller percentage of school funding comes from the national government. A relevant milestone document in the United States was the 1966 Coleman Report, which stated that "schools bring little influence to bear on a child's achievement that is independent of his [sic] background and general social context" (Coleman 1966, p. 325). It is widely cited in policy proposals, arguing that if many students performed poorly because of their respective backgrounds rather than because of the schools themselves, what was needed was not greater funding per se.

In the context of developing countries, the arguments are certainly framed differently, but questions remain on whether additional educational funding is worthwhile. Hanushek (1995, 1996), for example, has written on the poor correlation between school resources and student achievement in both industrialized and developing countries. Studies have also attempted to measure the

degree of local capture, or bureaucratic or political expenditures of school funds, across countries.

At first glance, studies such as those by Hanushek do appear to show little or no correlation between school resources and student achievement. Some of these studies have also been criticized on several fronts. First, some contain statistical flaws, such as the lack of longitudinal data, which render the results less robust. Second, they lack nuance when they use only aggregate data, since an overall low correlation may actually mask large, potentially meaningful, fluctuations. Third, these studies assume that measures of student achievement are accurate, and that assertion remains controversial. Fourth, standardized score studies fail to account for the consequences incurred by pressure to attain high score gains, such as increased difficulties in teacher hiring and disproportionate effects upon minority retention rates. Finally, questions can be raised regarding the validity of student performance measures; large debates surround the standards movement overall (and are discussed in greater detail later in this entry).

Such studies are therefore more likely to mobilize pre-made opinions than to make significant contributions to policymakers. The appropriate question, then, is not, *Does money matter?* Instead, it is, *When does money matter, and how?*

As these cases suggest, educational funding is controversial partly because there are few experimental studies in schooling, the way there are in medical trials. Good data remain scarce. Still, one major randomized, experimental study on class size measured achievement test results from Project STAR in Tennessee and showed that small class size does matter, that its impact remains even after students return to larger classes in later grades, and that the effects were larger on

minority and poor students (Krueger 1999). Another study followed 123 African Americans from similarly low-income households in Chicago for over 40 years. Approximately half of them were placed in preschool, and the remaining half constituted a control group. Results show that while second-grade test scores for the two cohorts appeared to be similar, those who attended preschool report higher incomes, lower crime rates, higher marriage rates, and educational attainment (Schweinhart et al. 2004). A similar, larger-scale program in the United Kingdom, called Sure Start, was launched in 2000 and thus far also shows limited but significant benefits to early schooling interventions. A RAND study in California found that every dollar invested in pre-school programs produces a return of \$2 to \$4 (Karloly and Bigelow 2005).

One major policy lesson for developing contexts is many policy proposals are non-transferable, and like all educational policies, data must be disaggregated and compared to appropriate counterfactuals. For example, by pointing to radio education (in the form of subject-specific textbooks and corresponding lessons aired as radio programs) and textbooks as worthwhile expenditures, disaggregated studies refuted Hanushek's claim that money on school inputs does not affect school quality in developing countries (Kremer 1995; Hanushek 1995).

Structuring Education:

Private Versus Public Sectors

The question of how much government funding is enough is intrinsically tied to the politics surrounding the government's role in structuring education overall. Education is, for the most part, seen as a public sector activity for a number of reasons, including: an educated populace is considered a public good, schools are centers of citizenship training and public spheres of interaction, and

school systems are difficult to organize and can be characterized as natural monopolies, with large economies of scale.

Contemporary policies that question national governmental control over education are generally discussed under the umbrella rubric of "school choice". Four private-versus-public sector policies are prominent and widely debated: privatization of school administration, school vouchers, charter schools, and small schools. Because these policies are more likely to be overt in well-developed school systems, they are primarily discussed in the context of industrialized countries, particularly those that wish to introduce or deepen free-market policies. Therefore, most of the case studies lie in the United States, though school vouchers are a notable exception. Finally, charter schools and small schools are discussed as school reform policies aimed at urban areas.

Privatization of School Administration

Privatization of school administration occurs when government administrations award private firms contracts to run public schools. These firms receive public funds and in return, employ teachers, buy school materials, and care for all school activities. Contracts may come with curricular requirements and other provisions. Proponents of such privatization argue that private firms have greater incentives to run schools efficiently in order to save money and reap profits; opponents argue that such privatization does not come with enough quality assurance and accountability, that long-term activities such as education cannot be run by firms that can enter and leave the market at any time, and that the policy does not sufficiently reward pedagogical knowledge and experience.

Vouchers

Vouchers are payments made by the government to the private school of the

parents' choice, in lieu of funding to the public school the child would have otherwise attended. Voucher programs generally do not allow recipients to enroll in other public schools.

Particularly contentious in the United States is the policy of including parochial schools in a voucher system. Some argue that this policy, which in effect transfers public funds to religious institutions, violates the doctrine of separation between church and state. In some European countries, most notably the Netherlands and France, a policy akin to the voucher system has been long established. So-called "special schools" are those that receive some funds from the state, allotted according to the number of students in attendance, and must abide by regulations set by state boards. In France, this system of vouchers is called *école libre*, or "free schooling." In the 1980s, when it appeared that the system might be changed, mass demonstrations took place in protest. In the Netherlands, approximately 70% of children attend private, mainly religious special schools. There, that explicitly Catholic and Protestant schools receive public funds is considered uncontroversial and a reflection of religious diversity (Hirsch 1994).

Since the 1980s, the United Kingdom, New Zealand, Sweden, and Australia have also introduced vouchers or other public subsidies in order to encourage the growth of private sector schools. In Chile, a *de facto* voucher system was introduced in 1980 when private schools received government funds alongside public schools, according to the average number of students in attendance. Approximately one-third of Chilean children attend such private schools. As with the known European and American cases, the data show mixed results. McEwan and Carnoy (2000) found that non-religious voucher schools were marginally less effective, while religious voucher schools

were more cost-effective because they paid teachers lower salaries. In the United States, while it appears that many private schools operate with lower costs than public schools, and that voucher recipients are associated with better school performance than other students, the results are not significant when researchers control for selection bias among voucher recipients (Hoxby 1994).

Voucher proponents argue that vouchers can still help low-income and disadvantaged students have a better chance at affording and attending good schools, and that better outreach and distribution is needed for vouchers. Other voucher-supporters are community groups who view school choice not as a market issue but as a democratic one where one monolithic formula does not fit all (Hess, 1999); school choice overall allows for a diversity of schools reflective of societal pluralism (Fuller 2003).

Voucher opponents argue that vouchers do not necessarily enable low-income students to attend better schools, as many private schools administer entrance exams and admit only better-performing students. Neither do vouchers necessarily lead to better performing public schools, as taking public funds away may worsen their situation.

Charter Schools

Charter schools are viewed as a popular political compromise between typical public schools and vouchers because they operate within the public school system. As their names suggest, charter schools are guided by a pedagogical philosophy, curricular theme, or ideology described in their charter. The first ones were created in the American state of Minnesota in 1991; since then, over 80% of American states have written charter school provisions into their laws. In the United States, charter schools must abide by some of public school regulations, especially those dealing with fair admissions and

separation of church and state, but they are largely exempt from regulations dealing with curriculum and personnel.

As with vouchers, the results regarding charter schools are mixed. Many of them shut down when the founding director leaves, experience growing pains, or have financial troubles after a few years, while others flourish as small, distinct communities of their own. In a 2004 Department of Education study comparing test results for fourth-graders nationwide, charter school students performed worse in mathematics and reading than regular public school students did, but researchers warn that no sweeping conclusions can be made upon these reported results (US Department of Education 2004).

Small Schools

Finally, a very recent but burgeoning trend in school choice is that of “small schools.” These small schools are intended to provide students with the familiarity and attention of small communities, but within a public school system. Because these small schools come with their own funding (provided almost exclusively by the Bill and Melinda Gates Foundation) and abide by public school regulations, they avoid most of the controversy surrounding vouchers. Nevertheless, because few small schools are created each year, and these schools are often physically housed in the same overcrowded buildings attending students were hoping to avoid, small schools can also be critiqued for exacerbating a two-tiered system, thus aggravating tensions between regular public school and small school students. Policymakers therefore argue that in order to succeed, separate physical spaces are needed for these schools.

Student Performance & Standards Reform

The trends of school choice and privatization are inextricably linked to a growing emphasis

on measured student performance and standards reform. For all of these policies, proponents rely on cost data and standardized examination data, such as fourth-grade mathematics test scores, to determine whether schools are running efficiently. In turn, based on known standardized exam data, so-called efficient charter schools and public schools are rewarded with more funding and non-efficient ones face the threat of closure; and parents are encouraged to utilize vouchers to transfer their children to more efficient schools, whether public or private.

In the United States, the “No Child Left Behind” (NCLB) Act of 2002 rendered financial censures and rewards, based on standardized test scores and national policy. Supporters of the legislation argue that it increases *accountability* and compels teachers to ensure that students, even previously low-achieving ones, reach minimum standards. While most policymakers supported the legislation’s premise, popular support for the NCLB has eroded because of implementation problems. Critics argue that criteria used to determine Adequate Yearly Progress (AYP) remain ambiguous, that teachers fear censure and consequently feel compelled to “teach to the test” rather than focus on substantive content, and that in extreme cases, the legislation actually creates incentives for teachers to fail out students they think will lower test score averages, or to leave challenging low-income school districts for wealthier ones. All of these critiques are accompanied by an overarching complaint that NCLB is in effect an “unfunded mandate”, in that states are required to meet and enforce standards without the resources needed to do so. In turn, several states have begun to challenge the legislation’s legality in the nation’s judicial courts. Some critics also assert that participatory governance offers a better means of accountability than NCLB’s system of financial awards and penalties.

The No Child Left Behind Act is significant for reasons beyond its consequences on American schools. It is indicative of an overall emphasis on standardized test scores shared by the World Bank and many international development organizations. For example, a combination “expenditure tracking and service delivery” (ESDS) survey was conducted in Zambia in 2002, and private and public funds to schools are measured against test scores. To the extent that educational outcomes and notions of accountability are likely to be expressed via quantitative data, some current policy debates pivot on the validity of such outcome measures in the first place. Ideally, the standards reform movement creates concrete, achievable goals for school administrators, teachers, and students. However, critics contend that like all quantitative data, standardized test scores are easy to manipulate. For example, critics question how statistically significant trends are determined, as one year’s fourth-grade cohort can be very different from the next year’s, especially since many students move and change school districts between school years. Critics also question how schools disproportionately serving low-income or immigrant students should be judged. Finally, many question the construct validity of the exams themselves, arguing that even if a perfect fourth-grade test were somehow designed and conducted, standardized exams inherently exclude more nuanced evaluations of essays, problem solving, creative writing, political participation, community, and other goals of public education.

The Chilean case on vouchers mentioned earlier in this entry, then, is significant not only as an example of the tension between private versus public sectors, but also because its study was only possible alongside well-run databases of educational outcomes. As developing countries hire teachers and build

more schools and school systems to increase overall access to education, they also face increasing pressure to present educational outcomes other than general enrollment, meet universal criteria, and face economic sanctions or rewards based on these criteria.

Emergency Education

Emergency education consists of schooling conducted without stable governmental support and physical infrastructure or administration, often in the context of war, displacement, and the need of overall emergency humanitarian aid as well as traditional school activities. While refugee relief efforts, such as those for Chechnyans in neighboring republics or Myanmar refugees in Thailand, have focused on food, medical care, and mental health counseling, the role of emergency education in psychosocial healing has grown in prominence. Specifically, emergency education has proven to be pivotal in helping youth to engage in age-appropriate activities, restore a sense of “connectedness” with others, perceive a positive opportunity for leadership in the family, and support community- rather than individual-level development programs (Crisp, Talbot et al 2001). While literacy and numeracy remain important goals, especially in longer-term situations, the psychosocial aspects of schooling take precedence in emergency education. For the most part, partly because it usually transpires in the context of displacement or state conflict, emergency education appears to be primarily provided by *non-government organizations* rather than governmental agencies.

Major challenges in emergency education, besides its immediate provision in changing contexts, include the role of peace and *human rights* education in emergency schools, inter-agency coordination and governance, the need for increased parental and community involvement and governance, and the need

for better documentation of efforts (Education in Emergencies 2005).

Bilingual Education

For immigrants, longer-term refugees, citizens of postcolonial countries, and indeed, all residents of countries with significant immigrant populations, the policy of bilingual education is meant to facilitate literacy in the country's official language and in the student's native or indigenous language. For example, in the United States, classes taught in both Spanish and English are meant to help Latino immigrants learn English as well as retain Spanish. In contexts of high immigration and mobility, proponents of bilingual education also argue that native-English speakers also benefit from bilingual education, since traditional foreign language programs are limited. Although such programs primarily serve immigrant children, they may gain political support if more bilingual programs are seen as innovative programs for all children, rather than as programs catering to the special needs of immigrant children.

Because meta-analyses of results do not conclusively show that bilingual programs either hinder or facilitate English-learning among immigrants (Greene 1998), debates surrounding bilingual education tend to reflect those surrounding immigrant rights overall. That is, opponents to bilingual education programs argue that English-only schooling is necessary for immigrant children to become full citizens, while proponents of bilingual education argue that bilingual education not only enhances learning in both languages, it is integral to both democratic self-determination and a pluralistic society.

In postcolonial countries, recent efforts have been made to re-introduce indigenous language instruction into public schools. In some nations within Africa, for example, most schools conduct classes in English,

French, and Portuguese, languages of the former colonial powers; in Latin American countries with significant indigenous populations such as Bolivia and Guatemala, class is usually held in Spanish. In response, international non-government organizations encourage bilingual education programs. UNESCO's official policy holds that education in the mother tongue, listed in its 2001 Universal Declaration on Cultural Diversity, improves the quality of education, social and gender equality, inter-cultural education within and between societies. Nevertheless, implementation of such recommended policies vary; for the most part, bilingual programs remain pilot projects. In some countries with diverse linguistic minorities, the government has openly defied recommendations for multi-lingual education. In Burma-Myanmar, for example, the government has banned the languages of all ethnic minorities and forcibly implemented Burmese as the only sanctioned language of instruction.

Conclusion

Because of the importance of education in human capital development, especially as espoused by prominent international institutions such as the World Bank, current debates in education policy and schools tend to center on economics and financing. Thus, returns to education are often measured in increased national productivity or individual income, and schools face growing pressure to measure effectiveness via quantitative outcomes such as standardized student test scores. Partly because returns to education are long-term, it is often difficult to secure adequate funding for schools, especially those serving disproportionately marginalized populations such as girls, racial or ethnic minorities, or immigrant populations.

Even governance questions often center on funding structures of school systems; namely,

policymakers argue over whether schools are best run in private or public sectors. Within public sector school systems, proponents of participatory governance argue that decentralization of policy-making powers can lead to greater legitimacy and effectiveness of school policies, but only when local communities are given the administrative training, fiscal resources, and political support to properly implement their decisions.

Schools also continue to be centers of citizenship as well as incubators of human capital. Public policy debates surrounding schools and citizenship are especially contentious in times of political transition, in cases of emergency education, in countries with undemocratic governments, and in countries with large immigrant populations or linguistic diversity.

Selected References

- Armbruster, Ralph; Kim Geron and Edna Bonacich. (1995) "The Assault on California's Latino Immigrants: The Politics of Proposition 187", *International Journal of Urban and Regional Research*, 19, 4, 655-663.
- Betancourt, Teresa; Rebecca Stichick and Gillian Dunn. (2004) "The IRC's Emergency Education Programme for Chechen children and Adolescents", *Forced Migration Review*, 15, 28-30.
- Bourdieu, Pierre. (1986) "The Forms of Capital", in John Richardson (Editor), *Handbook of Theory and Research for the Sociology of Education*. New York: Greenwood Press, 241-258.
- Bowles, Samuel and Herbert Gintis. (1976) *Schooling in Capitalist America*. New York: Basic Books.
- Bowles, Samuel and Herbert Gintis. (1980) "Contradiction and Reproduction in Educational Theory", in L. Barton; R. Meigham and S. Walker, (Editors), *Schooling, Ideology, and the Curriculum*. London: Falmer Press, 51-65.
- Chimombo, Joseph. (2001) "Educational Innovations in Developing Countries: Implications and Challenges for Policy Change in Malawi", *Journal of International Cooperation in Education*, 4, 2, 39-54.
- Choi, Soo-Hyang. (2004) "Access, Public Investment, and Equity in ECCE", *UNESCO Policy Briefs on Early Childhood*. Paris: UNESCO.
- Coleman, James. (1966) *Equality of Educational Opportunity*. Washington DC: U.S. Government Printing Office.
- Crisp, Jeff; Christopher Talbot and Diana B. Cipollone. (2001) (Editors) *Learning for a Future: Refugee Education in Developing Countries*. Geneva: UNHCR.
- Easterly, William. (2001) *The Elusive Quest for Growth*. Cambridge, MA.: MIT Press.
- "Education in Emergencies". (2005) Special issue of *Forced Migration Review*, 22.
- Freeman, Richard B.. (1976) *The Overeducated American*, Academic Press.
- Freire, Paulo. (1970) *Pedagogy of the Oppressed*. (Pedagogia dos Oprimidos), New York: Continuum Books.
- Fuller, Bruce. (2003) "Education Policy Under Cultural Pluralism", *Educational Researcher*, 32, 9, 15-24.
- Fung, Archon. (2004) *Empowered Participation: Reinventing Urban Democracy*. Princeton, NJ.: Princeton University Press.
- Fung, Archon and Erik Olin Wright. (2001) "Deepening Democracy: Innovations in Empowered Participatory Governance", *Politics & Society*, 29, 1, 5-41.
- Grant Lewis, Suzanne and Shireen Motala. (2004) "Educational De/Centralization and the Quest for Equity, Democracy and Quality", in Chisholm, Linda (Editor), *Education and Social Change in South Africa Since 1994*. Cape Town: Human

- Sciences Research Council Publishers and Zed Press.
- Grant Lewis, Suzanne, and Jordan Naidoo. (2004) "Whose Theory of Participation? School Governance Policy and Practice in South Africa", *Current Issues in Comparative Education*, 6, 2.
- Greene, Jay P. (1998) *A Meta-Analysis of the Effectiveness of Bilingual Education*. Los Angeles: USC, Tomas Rivera Policy Institute.
- Hanushek, Eric. (1995) "Interpreting Recent Research on Schooling in Developing Countries", *World Bank Research Observer*, 10, 2, 227-46.
- Hanushek, Eric. (1996) "School Resources and Student Performance", in Gary Burtless (Editor), *Does Money Matter?* Washington DC: Brookings Institution.
- Hess, G. Alfred. (1999) "Community Participation or Control? From New York to Chicago", *Theory into Practice*, 38, 216-24.
- Hirsch, Donald. (1994) "Schools: A Matter of Choice", *OECD Observer*, 187.
- Hoxby, Caroline Minter. (1994) *Do Private Schools Provide Competition for Public Schools?* National Bureau of Economic Research, Working Paper. Cambridge, MA.: NBER.
- Karoly, Lynn A., and James H. Bigelow. (2005) *The Economics of Investing in Universal Preschool Education in California*. Santa Monica, CA.: RAND Institute.
- Knodel, John E. (2003) "The Closing of the Gender Gap in Schooling: The Case of Thailand", in E.R. Beauchamp (Editor), *Comparative Education Reader*. New York and London: Routledge, 183-215.
- Kremer, Michael. (1995) "Research on Schooling: What We Know and What We Don't: A Comment on Hanushek", *World Bank Research Observer*, 10, 2, 247-54.
- Krueger, Alan. (1999) "Experimental Estimates of Education Production Functions", *Quarterly Journal of Economics*, 114, 2, 497-532.
- Levy, Frank and Richard Murnane. (1992) "U.S. Earnings Levels and Earnings Inequality: A Review of Recent Trends and Proposed Explanations", *Journal of Economic Literature*, 30, 1333-81.
- Makuwira, Jonathan. (2004) "Non-Governmental Organizations. (NGOs) and Participatory Development in Basic Education in Malawi", *Current Issues in Comparative Education*, 6, 2.
- Mankiw, Gregory. (1995) "The Growth of Nations", *Brookings Papers on Economic Activity*, 1, 275-326.
- McEwan, Patrick and Martin Carnoy. (2000) "The Effectiveness and Efficiency of Private Schools in Chile's Voucher System", *Educational Evaluation and Policy Analysis*, 22, 3, 213-39.
- McLean, Gary N.; Kenneth R. Bartlett and Eunsang Cho. (2003) "Human Resource Development as National Policy: Republic of Korea and New Zealand", *Pacific-Asian Education Journal*, 15, 1, 41-59.
- Pini, M. and S. Cigliutti. (1999) "Participatory Reforms and Democracy: the case of Argentina", *Theory into Practice*, 38, 4, 196-202.
- Pritchett, Lant. (1999) *Where Has All the Education Gone?* Washington DC: World Bank.
- Rotberg, Iris. (2004) (Editor) *Balancing Change and Tradition in Global Education Reform*. Lanham, MD: Scarecrow Press.
- Schweinhart, L.J.; J. Montie; Z. Xiang; W.S. Barnett; C.R. Belfield and M. Nores. (2004) *Lifetime Effects: The High/Scope Perry Preschool Study Through Age 40*. Ypsilanti, MI.: High/Scope Press.
- Tendler, Judith. (2002) "The Fear of Education", in *Background paper for*

Inequality and the State in Latin America and the Caribbean. Washington DC: World Bank.

Tremblay, Karine. (2005) "Academic Mobility and Immigration", *Journal of Studies in International Education*, 9, 196-228.

UNESCO. (2005) *Financing Education: Investments and Returns*. Montreal, Canada: UNESCO Institute for Statistics.

U.S. Department of Education. (2004) *America's Charter Schools: Results from the NAEP 2003 Pilot Study*. Washington, DC.: National Center for Education Statistics 2005-456.

Vavrus, Frances. (2003) "Uncoupling the articulation between girls' education and tradition in Tanzania", *Gender and Education*, 14, 4, 367-89.

Vavrus, Frances and Lisa Ann Richey. (2003) "Women and Development: Rethinking Policy and Reconceptualizing Practice", *Women's Studies Quarterly*, Fall/Winter, 6-18.

Websites

UNESCO. www.unesco.org

World Bank. www.worldbank.org/education

Celina Su
Department of Political Science
Brooklyn College
City University of New York
Brooklyn, New York, USA
celinas@brooklyn.cuny.edu

Education Policy: Universities

James J.F. Forest

Introduction

University policymakers worldwide are engaged in a constant struggle to balance appropriate levels of public funding and control with the needs of society. Most of the major policy decisions in higher education fall into at least one of the following three categories: (1) providing adequate levels of student access to universities; (2) managing the funding issues associated with the provision and expansion of higher education; and (3) ensuring institutional and programmatic quality. These dimensions of concern dominate most contemporary governance and policymaking debates, and frame the current challenges of globalization faced by higher education.

Access and Admissions Policies

Responsibility for establishing university admissions policies varies across countries, but there are three most common approaches: national direction, institutional autonomy, and open admissions policies. In the first approach, the national government maintains a limit on the number of students to be admitted and is directly involved in the selection procedure of those enrolled. In Greece, for example, admissions policies and the selection of students for postsecondary study are determined directly at a national level. In fact, the national government plays a prominent role in formulating policies regarding access to higher education in most countries. A variety of factors influence these decisions, including social equity and national workforce needs. For almost all countries, the basic requirement for access to higher education is the successful completion of secondary education (e.g., a high school diploma or GED in the United States).

Additional selection policies and procedures are often imposed for certain programs, disciplines or institutions.

Other countries allow individual institutions to determine their own selection and enrollment policies. For example, universities and colleges in the United Kingdom and United States are viewed as autonomous, able to pursue their own admissions policies. The same is true throughout Scandinavia, Spain, Portugal, and Ireland, although in the latter case the application process is facilitated by a Central Applications Office. In these countries, policymakers may provide certain guidelines or general criteria, but universities here are free to decide for themselves how to fill their enrollment capacity. Many strive to ensure that students are admitted to postsecondary study who are adequately prepared and likely to succeed in college. Naturally, admissions policies vary widely between institutions, and a university may have separate admissions policies for several degree programs.

Finally, in a number of countries only the completion of upper secondary education (e.g., a high school diploma or G.E.D. in the U.S.) is required for admission to most programs of study at public institutions. For example, students in French-speaking countries of sub-Saharan Africa, who earn an upper general secondary school-leaving certificate (known as the baccalauréat) or its equivalent are guaranteed admission to a university. Similar policies exist throughout the Arab World and in Belgium, where there are very strong traditions of free access to higher education. With few exceptions, access requirements for universities in Central and Eastern Europe include the upper secondary school leaving certificate and a passing grade on entrance examinations. In Austria, most universities are legally obliged to admit all students who register, although the *Fachhochschulen* and some academies are

more selective. In Italy, the universities decide which faculties will offer either open or limited access, while in the Netherlands, universities have open access in principle, although the number of students admitted can be limited at the national level when the qualified applicant pool exceeds the nation's labor market needs.

In the United States, primary responsibility for providing access to higher education rests with individual states and their public systems of colleges and universities. The federal government's role is largely limited to the provision of financial aid, although beginning in the mid-20th century Congress began enacting legislation to promote enrollment and encourage university compliance on a number of social and equity issues. Prominent examples include the Servicemen's Readjustment Act (otherwise known as the GI Bill of 1944), which rewarded veterans of World War II by providing a free college education, and the 1947 Truman Commission Report, which publicly promoted universal access and encouraged states to develop a community college system across the nation. By 1965, over 50 percent of the nation's high school graduates were bound for college. The Higher Education Act (1965) and subsequent reauthorizations further articulated the scope of federal government higher education policy in the U.S., while the nation's fiscal and economic policies support a healthy climate for private institutions (both non-profit and for-profit) and distance learning programs. Overall, through an impressive array of public policies, the United States has for decades led the world in providing access to postsecondary study. During the 2002-03 academic year, nearly 15 million students were enrolled in a U.S. institution of higher education (U.S. Department of Education 2003).

The latter half of the twentieth century witnessed a dramatic rise worldwide in the demand for access to higher education, bringing considerable pressure on government and institutional policymakers. Within the last decade, several countries have developed and implemented strategies to significantly increase student enrollments through a mix of new public and private institutions, as well as open and distance learning initiatives. For example, 32 new universities have been established throughout the Arab World since 2000. In addition, hundreds of community colleges, technical higher education institutes, and colleges of technology have been established in this region, with considerable concentration in Lebanon, Egypt, Oman, Sudan, the United Arab Emirates, and the Palestinian Territories.

Policies that encourage private higher education have especially contributed to expanded enrollment in many regions, particularly in Asia, Central and Eastern Europe, and Latin America. In many countries of these regions, private colleges and universities now outnumber public institutions. Within the last five years, the number of private colleges and universities in Malaysia has increased from around 100 to 690, while in Bangladesh almost 100 new private higher education institutions were established between 1998 and 2001. Over the same period, 46 new private institutions were established in Mongolia, 20 in Nepal, and 11 in Costa Rica. Within the last decade, 18 new private universities have been established in Paraguay, and in Kazakhstan, the number of private higher education institutions increased from 41 to 123 (UNESCO 2003).

Distance learning initiatives are also playing a vital role in providing access to higher education. From extension courses to full-scale degree programs and online institutions, distance education opportunities

are particularly common throughout Asia and the Pacific (with 15 digital universities in Korea and 67 online colleges established by conventional universities in China), Latin America (particularly Brazil, where 80,000 students were enrolled in distance courses in the 2000-01 academic year), Africa and the Arab World. Prominent examples in Africa include Zimbabwe Open University, the Open University of Tanzania, and the African Virtual University, all of which were established after 1997. The National Open University of Nigeria, which was closed in 1985, re-opened its doors in May 2003 with 100,000 students attending 18 campuses distributed throughout the country. The Arab Open University was established in 1999 with its headquarters in Kuwait and branches in Bahrain, Egypt, Lebanon, Jordan and Saudi Arabia. The Syrian Virtual University, established in 2002, offers internationally accredited degrees, and has recently concluded agreements with western online universities – particularly in Canada and the United States – to provide their programs to Arab students.

Funding

While these and other expansion initiatives require considerable resources, even the wealthiest governments are unable to meet the increasing demand for higher education using exclusively public funds. Indeed, the funding of higher education has been a perennial governance and public policy dilemma, as policymakers worldwide are challenged to devise new strategies for reducing costs, increasing efficiencies and finding alternative funding streams. In recent decades, much of the debate over university funding has concerned the distinction between higher education as a public good versus a private good. Critics of public spending argue that university credentials are largely a private good, and thus public monies

should not be used to subsidize the costs of attending a university for a small minority of the nation's citizens. Supporters of broad access to university education contend that in a sophisticated service economy, higher education provides a public good (i.e., an educated workforce) that benefits the nation's competitiveness in the global marketplace.

To its critics, higher education is a costly, service-oriented industry in which costs are skyrocketing out of control. Some have argued that reducing public funding will not only make institutions more efficient, but more accountable to state demands. Thus, new models of resource generation and allocation – coupled with the desire for greater accountability – are being explored. Serious discussions have taken place throughout Europe, the United States and parts of Asia about market forces, performance-based models of resource allocation, deregulation of institutions and programs, and promoting greater efficiency. With an increasingly prevalent view that university education generates substantial personal benefits for its recipients, leaders of many developed countries have sought to transfer more of the costs to the students, often through some combination of tuition fees, student loans, or special taxes on graduates' income. These approaches are a marked departure from traditional means of public financing for higher education.

Generally speaking, there are two main methods of state financing for public universities: either the institutions are funded entirely by the public authorities, or the universities (at least those in the public sector) receive a state grant but at the same time charge registration or tuition fees. Block grants, except for targeted research funding, are common in many countries, including Australia, Denmark, the Netherlands, Canada, Sweden and the United Kingdom. University education is provided free of charge for

students in Denmark, Greece, Luxembourg, the Czech Republic, Hungary, and Poland. In the Nordic countries, Bulgaria, and Malta, students pay no tuition fees and receive additional support to cover the cost of living. Public institutions in Iceland have modest registration fees, but are otherwise considered free. In England, Wales and Northern Ireland, students contribute towards their tuition fees (up to a quarter of the full cost) based on a means test, and around 50% of students are not required to make any contributions. In Germany, Greece, the Czech Republic, Romania and Slovakia, additional financial support is awarded to the parents of students, including payments and tax relief.

In Estonia, Latvia, Lithuania and Romania, most students pay tuition fees, while in several other countries, students pay a mandated contribution (in addition to registration and/or tuition fees) to an organization other than the university: in France, a payment to cover medical care; in Italy, a tax paid to the regional bodies which administer all the forms of student support; in Austria and Iceland, a membership fee to the student organizations. In the United States, the tremendous diversity of higher education institutions and sectors translates to a broad spectrum of financial arrangements. At the community college level, registration and tuition fees are modest by any standards, while at many private research universities and liberal arts colleges students and their families can expect a bill for well over \$30,000. Scattered throughout the middle of this spectrum are hundreds of public universities and colleges, each funded at varying levels by their local state tax base (e.g., California, Illinois, New York). For both private and public higher education in the U.S., federal support is largely limited to research initiatives and student financial aid. Some countries have considerably decreased government support as a proportion of total

revenue for public higher education. One particularly interesting example is Australia, where universities now generate almost 50 percent of their total operating revenue through tuition fees charged to domestic and international students; external research grants; student contributions through the Higher Education Contribution Scheme (HECS); commercial activities; revenue from investments; and endowments and donations (UNESCO 2003). As a result, the Australian higher education system is now described officially by the Commonwealth Government as a “publicly subsidized” rather than a “publicly supported” system.

It is increasingly the case worldwide that students will be expected to pay some form of registration and tuition fees, and several types of financial assistance are typically made available to qualified students. This assistance can be targeted a particular groups of students (often students from financially or otherwise disadvantaged backgrounds) or can apply to virtually all students. Governments and institutions typically make available two primary forms of student financial aid: (1) grants and/or loans to cover tuition fees and the cost of living; and (2) allowances or tax credits for students and/or the parents of students. These kinds of financial support often complement each other, and a student’s financial aid package will typically contain a mix of grants and loans. Shifting the burden of finances to students raises a variety of equity and access issues, particularly in developing countries throughout Africa, Asia and Latin America. For example, interest rates on the repayment of student loans, and informal discrimination by institutions against students with loans, currently act as deterrents to many students. However, the World Bank requires its client countries in the developing world to adopt a “user pays” approach, as part of an overall strategy to reduce total public spending in these regions.

In sum, there is currently a worldwide movement towards privatization of higher education, reducing the university's financial dependence on public resources and shifting more of the burden to the students and their families. The outcomes of treating the university as similar to other private enterprises have not always been positive. For example, in New Zealand, the introduction of market forces into higher education has led to the bankruptcy of several higher-education institutions, a decrease in college enrollments from the country's poorest areas, and stratification of the elementary and secondary system by race and socioeconomic status (Courturier 2003).

Assessing and Ensuring Quality

Concurrent with these trends in funding policies, governments worldwide have shown a growing interest in measuring the outcomes of university activities. In many countries, funding decisions have become closely tied with assessments of performance quality, and policymakers in several corners of the world increasingly talk of 'return on investment' when discussing the public funding of higher education. In the U.S., many states now use some form of performance funding or budgeting system that directly links institutional outcome measures to financial resources. In the UK, the passage of the Education Reform Act of 1988 dramatically changed the funding and governance relationship between Parliament and higher education, creating a University Funding Council whose resource allocation decisions are based upon performance assessments of institutional teaching and research quality. In both the UK and the Netherlands, performance indicators have been used to rank university departments for the purpose of research funding, while in Finland, a new funding model implemented in 1997 incorporates a performance assessment

component that constituted 3% of the entire university budget (UNESCO 2003).

The demands for accountability are worldwide and complex. At the institutional level, university leaders are continually challenged by all their constituencies and stakeholders to demonstrate the quality of their activities, while in many countries the general public has grown increasingly skeptical that the nation's investment in higher education is doing more than providing a decent lifestyle for professors and the promise of a lucrative career credential for pampered, upper-class students. Thus, many governments have recently established or strengthened their national systems of management and assessment procedures to monitor performance of higher education institutions, and to ensure quality. Predictably, efforts to assess the quality of universities have led to student and faculty protests in many countries, including Argentina, Mexico and the United Kingdom. However, the strength of the worldwide "accountability movement" in higher education will undoubtedly remain one of the most important dimensions of university policymaking for the foreseeable future.

Assessment and quality assurance initiatives are most commonly supported by a perception that the traditional measures of institutional performance and effectiveness—such as peer review and market choice—are not sufficient indicators of institutional value. One increasingly popular strategy for ensuring institutional quality is accreditation—a review by an external organization, resulting in a formal confirmation that certain quality standards are being met by the university. While accreditation has been common in the United States for decades, there is no such tradition in most parts of the world. However, national and regional accreditation strategies form a

major component of university accountability initiatives worldwide.

During the 1990s, a number of countries in Central and Eastern Europe established accreditation agencies. In Australia, the Australian Universities Quality Agency was established as a joint federal-state government initiative, with responsibility for academic audits of both universities and those state agencies responsible for the accreditation of private providers. In Thailand, the National Education Act of 1999 established an Office of Education Standards and Evaluation and mandated that all higher education institutions must be evaluated every five years. Similar legislation was passed in many Latin American countries, creating Argentina's National Commission for Evaluation and Accreditation (CONEAU), Brazil's National Education Council and Higher Education Board, Chile's National Commission of Undergraduate Accreditation (CNAP), and Costa Rica's National Accreditation System for Higher Education.

While most countries in the Arab World have rich educational traditions and governmental agencies responsible for higher education, prior to the Arab Regional Conference on Higher Education (Beirut, 1998) Jordan was the only Arab country to have established an accreditation agency. The Beirut Conference opened a new era of interest in quality assurance in this region, culminating in a resolution to establish a regional quality assurance and accreditation program—under the auspices of the Association of the Arab Universities—and calling upon government leaders to establish similar mechanisms at the national level (UNESCO 2003).

Regional initiatives are taking hold in other parts of the world as well. In Africa, while several countries have established national commissions for higher education, responsible for accreditation and quality

assurance, multinational organizations such as the South African Development Community (SADC) and the Economic Community of West African States (ECOWAS) are serving a vital role in identifying common assessment indicators and standards for the recognition of studies and degrees. Similar activities are led by the Association of Caribbean Higher Education Administrators (ACHEA) for countries in that region. In Europe, almost all countries have established national agencies and systems for the assessment of quality in higher education, and many of these are working together to foster broad cooperation in the development of regional standards and procedures for quality assurance. In June 1999, the ministers of education of some 30 European countries signed the “Bologna Declaration” the aim of which is to establish a European perspective in higher education and adopt a standardized system of credentials and degree qualifications.

In addition to national and regional initiatives, several global organizations are furthering the university assessment and quality assurance movement. The UNESCO-CEPES project *Strategic Indicators for Higher Education in the 21st Century* provides opportunities for exploring both existing standards and performance indicators and for formulating proposals for the increase of their relevance (UNESCO 2003). In some cases, quality assurance mechanisms have been developed as part of World Bank or Asian Development Bank projects, or as integral components of aid projects from donor countries. And voluntary organizations such as the International Network of Quality Assurance Agencies in Higher Education (INQAAHE) are seeking to develop an effective international approach to quality assurance.

A wide range of performance indicators—quantitative measures that attempt

to assess the achievements of higher education institutions and systems—are employed by these initiatives, which seek to measure a variety of “inputs” and “outputs.” For the former, policymakers are concerned with the quality of courses, faculty credentials, standardized test scores of students admitted to the university, appropriate budgeting mechanisms and maintenance of the physical infrastructure (laboratories, buildings, and libraries), among other measures. These types of “input” indicators of quality are supported by assumptions about the basic necessities for effective teaching, learning and research. For example, it is generally agreed that a university which crams over 2,000 students in a lecture room designed for 700 fails to ensure a quality educational experience. By the same token, observers look poorly upon institutions who fail to hire adequate number of faculty to teach the students they admit. This is a particularly difficult problem in countries throughout the developing world that struggle with the global brain drain in attracting and retaining qualified faculty. For example, during the late 1990s nearly 40% of the university faculty positions in Ghana and more than 60% of those in polytechnics were vacant, while more than 50% of the faculty positions in Nigerian universities were vacant (UNESCO 2003).

Policies and initiatives toward ensuring the quality of universities cover more than just the effective management of programs and stewardship of resources. Indicators of “outputs” (like “research productivity” or “students employed within six months of graduation”) are also considered a vital component for demonstrating university quality. Graduation rates provide an important indicator of the effectiveness of higher education systems and of their specific institutions, as any increase in the dropout rate is a worrying indicator of an important

human and financial waste of public and private resources. Policymakers are also increasingly concerned with measuring student outcomes beyond mere graduation rates, encompassing a grander sense of indicators that show student cognitive development as a direct result of the university experience. Definitions of desired graduates’ competencies differ widely across countries and institutions. In one example, the Beirut Declaration states that universities should aim to produce self-aware, independent-thinking, responsible, skilled, qualified and professionally capable citizens who are able to meet societal needs while providing expertise, critical perspective and ethical direction on social development, science and technology.

Overall, there are a wide variety of views concerning the definition of quality in higher education and how to ensure it. At one end of the spectrum surrounding this debate are those who feel that what is needed is an internationally developed and agreed-upon set of standards for measuring the quality of all the world’s universities. The European Network of Quality Assurance Agencies (ENQA) is a prominent example of a multinational initiative to develop a framework for assessing quality in higher education. The Asia Pacific Quality Network and the ASEAN University Network are in the process of developing common quality criteria, appointing chief quality officers in member institutions, strengthening national statistics collection and analysis, and achieving an overall higher degree of compatibility between national sets of performance quality indicators (UNESCO 2003).

Opposed to the regional/global standardization perspective of these initiatives are those who feel that the cultural and geographic uniqueness of a nation’s higher education system, and its relationship to

meeting the needs of the local and regional society, defy broad standardization. Proponents of this view highlight the importance of measuring the relevance of a university education, and call for an examination of academic programs and degrees offered by institutions to ensure that the workforce needs of the country will be met by graduates in those programs. Leaders in many countries are particularly concerned with overcoming current shortages of specialists for the knowledge and information industries and ensuring that all graduates have at least basic competencies in the use of information technology. In Europe, the Bologna Declaration states as one of its main objectives the promotion of the employability of European graduates. Policymakers are also concerned with ensuring that degrees and diplomas awarded by their country's universities will be recognized by governments, universities, and employers in other countries. The issue of international recognition is becoming increasingly important in the age of multinational corporations and globalization.

Conclusion: University Policy and Globalization

The university is both local and global. Globalization is a phenomenon which significantly impacts higher education and to which universities and college contribute, particularly in the economic, technological, and scientific realms of activity. As international trends reveal, globalization impacts university policy in all three of the most important dimensions – funding, access, and quality assurance. While funding for higher education has traditionally been the responsibility of national governments, a worldwide shift toward privatization (particularly involving dramatic increases in the proportion of funding contributed by the students and their families) is having the

effect of most any other market; viewing students as clients transforms universities into an entity requiring increasingly sophisticated marketing and sales techniques to attract and retain loyal, paying customers. And as this higher education market has become increasingly international, students are seeking a quality return on their investment, even if this means attending a university in another part of the world. Asian students are by far the largest group of students studying abroad worldwide, largely because their national leaders have failed to provide adequate access to a quality university education at home.

According to studies by UNESCO (2003), at the turn of the millennium more than 1.6 million foreign students were enrolled for higher education studies in the 50 major host countries. More than three-quarters of all foreign study takes place in just ten host countries: United States (with more than 30 percent of all foreign students), France (more than 11 percent), Germany (about 10 percent), the United Kingdom (about 9 percent), the Russian Federation (about 5 percent), Japan (around 3.5 percent), Australia (about 3 percent), Canada (less than 2.5 percent), Belgium (less than 2.5 percent) and Switzerland (about 2 percent). Several regional policy initiatives are contributing to regional and international student mobility, including Europe's ERASMUS and SOCRATES programs in Europe, NAFTA and Mercosur in the Americas, and ASEAN and APEC in the Asian and Pacific regions.

Internationalization is also making inroads in the domains of quality assessment and accreditation. Issues of accountability (assessing the outcomes of a nation's colleges and universities) has become increasingly important throughout the world, as leaders struggle to harness costs while meeting the country's demands for access to an advanced degree. A prominent example is Universitas

21, headquartered at the University of Melbourne, which brings together a group of comparably-sized public universities from Australia, Canada, New Zealand, Singapore, the United Kingdom and the United States to accredit each other and share external examiners.

Overall, globalization compounds the challenges faced by university and national leaders in seeking to meet the increasing demand for access to higher education, maximize resources, and ensure the quality of their institutions and programs. Clearly, the ways in which policymakers respond to globalization may ultimately decide the fate of the modern university.

Selected References

- Altbach, Philip G. (2003) "Globalization and the University: Myths and Realities in an Unequal World", *Current Issues in Catholic Education*, 23, Winter, 5-25
- Altbach, Philip G. (1991) (Editor) *International Higher Education: An Encyclopedia*. 2 volumes. New York: Garland Press.
- Bleiklie, I. (2001) "Towards European Convergence of Higher Education Policy?", *Higher Education Management*, 13, 3, 9-30.
- Campbell, C. and van der Wende, M. (2000) *International Initiatives and Trends in Quality Assurance for European Higher Education*. Helsinki: European Network of Quality Assurance Agencies.
- Clark, Burton R. (1995) *Places of Inquiry: Research and Advanced Education in Modern Universities*. Los Angeles: University of California Press.
- Coururier, Lara K. (2003) "Balancing State Control with Society's Needs", *Chronicle of Higher Education*, 49, 42, p. B20 (June 27)
- De Witt, Hans. (2002) *Internationalization of Higher Education in the United States and Europe*. Westport, CT: Greenwood Press.
- Dill, David. (1995) "Through Deming's Eyes: A Cross-National Analysis of Quality Assurance Policies in Higher Education", *Quality in Higher Education*, 1, 2, 95-110
- El-Khawas, Elaine H. (1998) "Accreditation's Role in Quality Assurance in the United States", *Higher Education Management*, 10, 3, 43-56.
- European Center for Higher Education, UNESCO-CEPES. (2003) *Trends and Developments in Higher Education in Europe*. Paris: UNESCO
- Forest, James and Kevin Kinser. (2002) *Higher Education in the United States: An Encyclopedia*. 2 volumes. Santa Barbara: ABC-CLIO Publishing.
- Fulton, O. and Jurgen Enders. (2002) (Editors) *Higher Education in a Globalizing World: International Trends and Mutual Observations*. Dordrecht, the Netherlands: Kluwer Publishers.
- Goedegebuure, Leo; Frans Kaier, Peter Maasen, Lynn Meek, Frans van Vught, and Egbert de Weet. (1994) "International Perspectives on Trends and Issues in Higher Education Policy", in Leo Goedegebuure; Frans Kaiser; Peter Maasen; Lynn Meek; Frans van Vught; and Egbert de Weert (Editors), *Higher Education Policy: An International Comparative Perspective*. New York: Pergamon Press, 315-348
- International Association of Universities. (2003) *Internationalization of Higher Education: Trends and Developments since 1998*. Paris: UNESCO
- International Institute for Higher Education in Latin American and the Caribbean. (2003) *Reforms and Innovations in Higher Education*. Paris: UNESCO
- Neave, Guy and Frans van Vught. (1994) (Editors) *Government and Higher*

- Education Relationships Across Three Continents: Issues in Higher Education*. Tarrytown, NY: Elsevier Science.
- Organization for Economic Cooperation and Development. (2002) *Education at a Glance: OECD Indicators*. Paris: OECD.
- Robertson, Margaret. (1998) "Benchmarking Teaching Performance in Universities: Issues of Control, Policy, Theory and "Best Practice", in J. Forest (Editor), *University Teaching: International Perspectives*. New York: Garland Publishing.
- Teferra, Damtew and Philip G. Altbach. (2000) (Editors) *African Higher Education: An International Reference Handbook*. Bloomington: Indiana University Press.
- UNESCO Regional Bureau for Education in the Arab States. (2003) *Higher Education in the Arab Region*. Paris: UNESCO
- UNESCO Regional Bureau for Education, Dakar. (2003) *Recent Developments and Future Prospects of Higher Education in sub-Saharan Africa in the 21st Century*. Paris: UNESCO
- UNESCO Regional Office, Bangkok. (2003) *Higher Education in Asia and the Pacific, 1998-2003*. Paris: UNESCO
- World Bank. (2000) *Higher Education in Developing Countries: Peril and Promise*. Washington, DC: World Bank Task Force on Higher Education and Society.
- World Bank. (2002) *Constructing Knowledge Societies: New Challenges for Tertiary Education*. Washington, DC: World Bank.
- Boston College Center for International Higher Education. www.bc.edu/cihe
- United National Educational, Scientific and Cultural Organization (UNESCO). www.unesco.org
- World Bank. www.worldbank.org
- Institute of International Education. www.iie.org
- Organization for Economic Cooperation and Development. www.oecd.org

*James J.F. Forest
 Combatting Terrorism Centre
 United States Military Academy
 West Point, New York, USA*

Internet Sites

- Higher Education Resources Network. www.higher-ed.org
- International Comparative Higher Education Finance and Accessibility Project. www.gse.buffalo.edu/org/IntHigherEdFinance

Efficiency and Equity

John Davis

Introduction

Equity and efficiency strike most policy-makers and individuals as principles that are equally important in a global society and world economy. On the one hand, world economic development brings the greatest benefits to people if pursued as efficiently as possible, while on the other, fairness has been enshrined as one of the most fundamental human values. But until recently (e.g., Blank 2002; LeGrand 1990) most economists have seen efficiency and equity as conflicting normative principles that cannot be simultaneously satisfied, and which thus need to be traded-off against one another in the formulation of economic policy (e.g., Myles 1995). The view was put especially clearly by Brian Barry (1965) who argued that rather than suppose one value dominates others, they might be thought to be substitutable for one another. Barry used an equity-efficiency trade-off as his main example. In perhaps the most influential use of the idea, Arthur Okun (1975) explained the trade-off as between efficiency in the sense of economic growth and equity understood as equality, and regarded the tradeoff as inescapable. But the idea has also been applied in connection with such matters as income distribution (Sheshinski 1972), poverty and development (1996), the delivery and distribution of health care (Wagstaff 1991), and the financing of education (Hoxby 1996). Yet why trade-offs exist, and in what degree or manner trade-offs occur, has been complicated by debate over how to interpret the two principles.

For many years efficiency was seen to be relatively unproblematic (but see LeGrand 1991), and equity subject to competing interpretations. Indeed, the efficiency principle, which comes out of the nineteenth

century English utilitarian tradition of Jeremy Bentham (1970 [1789]), has enjoyed since the 1930s a generally agreed upon definition among economists in terms of the well-known Pareto criterion. A particular allocation of resources, either existing or planned, is defined as Pareto-efficient when it is impossible to re-allocate the resources concerned in such a manner as to make at least one person better off without at the same time making someone else worse off. Further, generally it is held that free markets are Pareto-efficient on the grounds that individuals can be said to pursue their own advantage in free markets. If individuals trade voluntarily, they must suppose they are as well-off as they can be. Thus in free markets it is impossible to re-allocate resources so as to make at least one person better off without making someone else worse off. (Economic efficiency or allocative efficiency as an economic policy criterion is thus different from technical efficiency or efficiency in production, i.e., that no more of an output can be produced from a given amount of input.)

In contrast, the principle of equity, which is usually associated with an egalitarian tradition of ideas linked to such individuals as Jean-Jacques Rousseau and Karl Marx, has long been defined in quite different ways. Sometimes it is treated as being synonymous with justice and fairness. Sometimes it is closely associated with equality between individuals, where this traditionally has involved either an equality of income or equality of utility, but more recently it has been seen to involve equality in individual capabilities to carry out basic activities of life. Yet another, recent interpretation of equity due to economists in the Walrasian tradition regards an equitable or fair allocation of resources as one which is 'envy-free,' that is, one for which no individual prefers any other individual's allocation to his or her own.

Let us put aside these differences in the way equity has been understood for the moment, however, to consider what is involved in the general idea of efficiency-equity tradeoffs. One can begin by imagining societies in which markets work freely and efficiently, but there is still considerable inequality, for example, inequality in wealth which makes it difficult for many individuals to participate successfully in markets. They may have so little capital or skills that they always find themselves limited to the most unattractive opportunities. In such societies, the efficiency principle operates, but many would say the equity principle does not, simply because it strikes most people as unfair that some can never escape the accident of their birth as less well-endowed individuals. But to remedy this situation typically involves limiting the market process—and thus efficiency. For example, were resources transferred to the less well-off to increase their prospects in the market system, say, through increased expenditure on public health and education, housing and food subsidies, etc., this would require moving resources from private production to the public sector through a system of taxation. Most economists would then argue that efficiency is lower despite the fact that equality has been enhanced through the benefits these types of social services provide the less well off. Moreover, since economists regard efficient allocations of resources through free markets as maximizing community income, trading-off efficiency for more equality would mean less to go around, or a smaller ‘pie,’ albeit with a more even distribution of that smaller ‘pie’.

This trade-off problem even arises when it comes to how governments raise taxes for possible social programs. Under usual assumptions employed in the theory of optimal taxation of commodities, it is efficient to place taxes on goods (sales or

excise) with low elasticities of demand, where low elasticity of demand means that the demand for these goods is unresponsive to price increases brought about by the addition of a tax. But goods whose demand is unresponsive to price changes are typically necessities such as food and housing. Optimal taxation would accordingly imply placing the highest taxes on necessities and the lowest taxes on luxuries (Ramsey 1927; also cf. Samuelson 1986), so that low-income households least able to bear those taxes would pay a disproportionately larger share of their income in taxes than high-income households. Thus even when social services for the less well-off are supported, doing so efficiently comes at the expense of equity.

Economists, then, generally believe that efficiency and equity are conflicting normative principles that cannot be simultaneously satisfied. This conclusion, however, is complicated by the fact that although the Pareto efficiency principle enjoys a generally agreed upon definition among economists, it has been increasingly criticized for over a half century as offering too narrow a foundation for economic policy. This critique has led to a closer examination of the Pareto definition, and a consequent recognition that it makes implicit use of number of non-utilitarian normative principles that have become central to recent theories of equity. To get better understanding of the efficiency-equity relationship, then, the second section of this entry begins with an examination of the history of the Pareto criterion, and elicits three non-utilitarian, non-efficiency normative principles implicit in it that play an important role in the most influential theory of equity in recent decades, namely, John Rawls’ influential account of justice as fairness. The second section then considers what Rawls’ account tells us about the efficiency-equity relationship, and argues that

the weight given to the principle of equality plays a key role in assessing efficiency-equity relationship. The third section, then, examines the current, leading theory of equity that also emphasizes equality, Amartya Sen's capabilities approach, and considers what this account additionally tells us about the efficiency-equity relationship. The fourth section takes a further look at the difficulties involved in explaining the principle of equality. The final section comments on the situation of policy-makers face when confronted with the question of the relationship between equity and efficiency.

Pareto Criterion and Rawls' Theory

It can be argued that an important stimulus to thinking about equity in economics were the problems that the New Welfare Economics encountered in attempting to restrict normative economics to the Pareto principle by eliminating the interpersonal comparisons of utility employed in the Old Welfare Economics of Alfred Marshall (1920 [1890]). When we make interpersonal comparisons of utility, we say we can compare how consumption of goods affects the utility or satisfaction of one person compared to how it affects the next. Marshall thus employed interpersonal comparisons of utility to identify states of affairs with the highest aggregate utility, and did so in a highly egalitarian way. Assuming that the utility of money used to buy goods decreases with income, aggregate utility could be increased by transferring income from high income to low income individuals until the point at which incomes were equal. This could be done through combinations of strongly progressive income tax systems and transfer programs. Then, were equity taken to mean equality, an efficient state of affairs, understood as the maximum sum of utility across individuals, was also an equitable state of affairs.

Against this, Lionel Robbins argued that different individuals' utilities could not be compared, and that interpersonal utility comparisons were impossible (1935 [1932]). This led to the New Welfare Economics abandonment of interpersonal utility comparisons and adoption of the Pareto criterion as the sole principle for evaluating alternative states of affairs. Yet in the absence of interpersonal comparisons the new approach lacked a means of determining which among different Pareto-optimal possibilities should be preferred. Thus a proposed extension of the Pareto criterion (Kaldor, 1939; Hicks, 1939) reintroduced interpersonal comparisons in non-utility terms, suggesting that one state of affairs could be regarded as preferred to another even when it made some individuals worse off if—in principle—the gains of those made better off were sufficient to compensate those made worse off, while leaving a surplus. There were technical difficulties with this proposal, but perhaps more important was the unacceptability of the idea that compensation only needed to be possible in principle. That a state of affairs could be recommended as preferred without compensation actually being paid simply seemed to violate fairness.

A second manifestation of the New Welfare Economics' difficulties deriving from the abandonment of interpersonal utility comparisons emerged in Kenneth Arrow's famous impossibility theorem which showed that on a number of standard assumptions about individual preferences—including that "interpersonal comparison of utilities has no meaning" (Arrow 1963 [1951]:9)—one was unable to construct a consistent criterion for social choice without assuming one individual's preferences to be dictatorial. Together with the Pareto criterion, this so narrowed the scope of application of the New Welfare Economics as to render it essentially meaningless for policy purposes. The

conclusion, then, that ultimately emerged was that Arrow was wrong to rule out interpersonal comparisons. Sen was particularly influential in arguing this (Sen, 1977). But interpersonal comparisons clearly raised justice and fairness issues. Thus the deeper lesson was that the New Welfare Economics had been unsuccessful in its attempt to construct a normative economics in terms of efficiency alone, and that the issue henceforth became how efficiency and equity were to be related.

In this regard, it helps to look more carefully at the Pareto principle to better understand its appeal as well as its indirect influence on subsequent thinking about equity. What can be seen is that the principle makes implicit use of non-utilitarian, non-efficiency principles that have become central to recent theories of equity. Three such principles particularly important to thinking about equity are the following.

(1)*Impartiality*. Because it sets aside distributional issues, the Pareto principle has often been taken to be relatively uncontroversial, so much so that it has even been thought to be ethically neutral and not a value judgment at all. Strictly speaking this is not accurate, since among the value judgments it conveys are that welfare-enhancing changes are desirable, that status quo arrangements are preferred, and that distributional issues can be set aside. Nonetheless, this apparent neutrality can be understood as a principle of impartiality, which is one dimension of equity or fairness.

(2)*Freedom*. The Pareto principle also has a close connection to the principle of freedom. Market exchange is generally thought to be mutually advantageous to exchangers on the grounds that otherwise they would not trade. Free markets are thus Pareto efficient; alternatively Pareto efficient states of affairs are free states of affairs. Thus

the value of freedom is implicit in the Pareto principle.

(3)*Respect for individuals*. The Pareto principle emphasizes respect for individual preferences. Indeed, since individual preferences are often not respectable (they may be malevolent, obnoxious, frivolous, etc.), an emphasis on the sanctity of individual preferences might better be seen as an emphasis on the importance of individual autonomy and a respect for individuals. Thus the principle of respect for individuals is also implicit in the Pareto principle.

That each of these three principles plays an important role in what has been arguably the most influential theory of equity in recent decades, namely Rawls' theory of justice (1971), clearly puts the idea of a trade-off between efficiency and equity in a different light.

Indeed Rawls saw his theory of justice in part as a response to New Welfare Economics, and employs all three of these additional normative principles above in arguments that apply directly to the relationship between efficiency and equity. Rawls' theory is most closely associated with the 'veil of ignorance,' the device he employs to establish the two principles of justice his theory introduces. We are asked to imagine what principles of justice individuals would adopt were they to decide from behind a 'veil of ignorance' regarding their own identities and individual characteristics. Rawls thus uses the 'veil' as a mechanism for insuring impartiality in order to claim that the principles of justice he will present are perceived to be fair. The first of these principles, moreover, is that individuals are to be entitled to as much freedom (or rights and liberties) as is compatible with all individuals having the same freedoms. The second principle (the maximin or 'difference principle') holds that social and economic inequalities are permitted if attached to

positions available to all, and arranged so as to be to the greatest benefit to the least advantaged members of society. We might interpret this principle as one meant to preserve the autonomy of individuals, both because of the protections it provides to the least well off, and because it preserves opportunity for different outcomes for different individuals.

Rawls' account might thus be said to draw out the implicit content of the Pareto principle in order to offer an alternative approach to how resources should be allocated between individuals. His first principle – that individuals are entitled to as much freedom as is compatible with all individuals having the *same* freedoms – requires an equal distribution of what he terms 'primary goods,' namely, "rights and liberties, powers and opportunities, income and wealth", and also "self-respect" (Rawls, 1971, p. 62). It thus constitutes a departure from Pareto allocations of goods by putting equity above efficiency, since achieving equality in primary goods would make some individuals worse off under any existing economic system. It is true that his second maximin or 'difference' principle modifies the extent to which this occurs by re-introducing inequalities if they benefit the least well-off, thus giving weight again to efficiency, but efficiency considerations are still confined within the framework of a prior concern for maximal freedom and equal distribution of 'primary goods'.

The question, then, is how significant this re-balancing of these two principles is in understanding how they are related. The answer depends upon how broad an understanding of 'primary goods' one employs, which itself is a matter of how what one believes to be necessary for individuals to have the *same* or equal freedoms. Rawls' characterization of 'primary goods' might include things ranging from the most basic

necessities such as food and shelter to things additionally regarded in many contemporary societies as necessities: education, health care, old age pensions, employment opportunities, safe communities, and so on. One might even argue that an ability to participate in a society's political processes would be considered an essential 'primary good' necessary to individuals having equal freedom. From this perspective, then, Rawls' approach is potentially quite open-ended, and thus his solution to the trade-off problem between efficiency and equity is to so broaden the understanding of equity as to define the space left for efficiency considerations. Rather than trade-off two independent principles, efficiency is defined in terms of equity.

Egalitarianism: Sen's Capability Analysis

Like Rawls, Sen's thinking developed in part as a response to what he perceived to be deficiencies in the New Welfare Economics' strong emphasis on the principle of efficiency and its view that efficiency needed to be traded-off against equity. Also like Rawls, Sen drew on the implicit content of the Pareto criterion in formulating his own views. But Sen rejected a central element in Rawls' account – namely, the latter's emphasis in his first principle of justice on equity requiring there be an equal distribution of "primary goods" – and developed his own approach in terms of equity requiring an equality of capabilities. In addition, his critique of the New Welfare Economics was deeper than Rawls' in that he built his case for equality of capabilities around the need to make use of interpersonal comparisons of capabilities between individuals – which the New Welfare Economics had rejected in the form of interpersonal utility comparisons, and about the need for which in any form Rawls had been ambivalent (Rawls 1971:90ff).

Sen draws on and re-develops the same implicit content of the Pareto criterion identified above in terms of the principles of impartiality, freedom, and respect for individuals. The principle that there should be an equality of capabilities is interpreted to mean that all should be equally able to develop the different capabilities that they have and value, not that all should develop the same capabilities. Calling for an equality of capabilities thus emphasizes the value of impartiality, since people are to be treated the same despite their differences. Calling for an equality of capabilities also emphasizes the value of freedom, since Sen sees individuals' development of their capabilities as "both a primary end and as the principal means of development", where the latter "consists of the removal of various types of unfreedoms that leave people with little choice and little opportunity of exercising their reasoned agency" (Sen, 1999, p. xii). And calling for an equality of capabilities emphasizes a respect for individuals in that it treats the individual as an agent in the "sense as someone who acts and brings about change, and whose achievements can be judged in terms of her own values and objectives" (*Ibid.*:19).

That an equality of capabilities requires that we make interpersonal comparisons between individuals is due to the fact that individuals differ markedly from one another in terms of their physical and mental capacities. Were they basically the same in these ways, one could pursue an equality of capabilities by seeing that all had essentially the same sorts of resources. Conversely, assuming individuals to be very different from one another in terms of their capabilities implies we must compare how they need different sorts of resources to equally achieve their respective capabilities. So making interpersonal comparisons is inescapable. This point is not readily apparent in the

framework of New Welfare Economics, since it sees interpersonal comparisons in utility terms, and the problematic nature of the idea that one might compare the subjective utilities of different individuals gets in the way of thinking about how we can generally compare objective differences between individuals in terms of their observable capacities. But Rawls's focus on an equality of "primary goods" also comes in for criticism in this regard. Since "primary goods" can be seen as simply basic resources, providing equal 'primary goods' does not ensure equality between people with different capabilities for transforming those same resources or "primary goods" into the sorts of achievements they value. A genuinely equitable world, then, would place greater weight on an equality of outcomes, where for Sen this is a matter of what capabilities people can actually acquire.

How does this all apply to the question of the relationship between efficiency and equity? Here, interestingly, Sen follows Rawls in important respects. Calling for an equality of capabilities—like calling for an equality of "primary goods" (Rawls' first principle of justice)—constitutes a departure from Pareto allocations of goods, and trades off efficiency for equity, since achieving equality in this regard would make some individuals worse off under any existing economic system. Similarly, Sen supposes that Rawls' difference principle (his second principle of justice), which modifies this trade-off by re-introducing inequalities if they benefit the least well-off, is not unreasonable. That is, one might allow inequalities in capabilities between individuals if this was to the advantage of those who were least well-off in terms of their capabilities development. Again, then, since the least well-off are not worse off, such inequalities would be Pareto efficient, and fairness in this respect would be

made somewhat more compatible with efficiency.

There are good reasons, however, to be hesitant over this accommodation of Sen's equivalent of Rawls' difference principle with efficiency (DeMartino 2000:109-110). When one moves from abstract theory to actual policy-making processes, how gains and losses resulting from policy changes actually materialize becomes an issue. Consider policies expected to be to the advantage of the least well-off in the future that depend on creating inequalities in capabilities in the short run. Should the expected benefits either not materialize or be less than anticipated, the real effect of the policy change is a worsening of the situation for the least well-off, and a state of affairs less fair than originally. Like the New Welfare Economics view of Kaldor and Hicks that all that is needed is that compensation be in principle possible, Sen's adoption of Rawls' difference principle may lead in practice to actual violations of fairness. Moreover, note that Rawls' 'veil of ignorance' device, which limits self-serving behavior, applies only to the foundational principles of justice and not to everyday policy-making. If we suppose that the least well-off are less likely to be as involved in policy-making as those who are better-off, then there is an additional risk that fairness will be violated by adoption of Rawls' difference principle.

A Further Look at Equality

A strong rationale for regarding equality as central to the understanding of equity is the idea that individuals somehow need to be in the 'same' position for a state of affairs to be equitable. Indeed, this simple idea underlies standard application of the equity principle in the economics of taxation (e.g., Musgrave 1959; Atkinson 1980), where horizontal equity – that individuals in the same position be treated equally – is distinguished from

vertical equity – that that individuals in different positions be treated differently. But in what respect are individuals to be treated the 'same?' Or as Sen puts it, what aspect(s) of a person should we regard as fundamental when we emphasize the importance of equality (Sen 1980).

We noted at the outset that some emphasize equality of incomes and some emphasize equality of utilities. The main problem with the former is that equal incomes can be unfair if individuals are thought to have different needs or vary in their desire for leisure. Equality of utilities, welfare, or satisfactions would then be the more inclusive standard, allowing one to balance out more income with greater needs and less income with a preference for leisure. A problem with utilities equality, however, is that utilities are difficult if not impossible to observe and measure in an objective manner. Another problem with utilities equality, drawn attention to by Rawls (1971), is that some individuals' satisfactions may be morally offensive. If the utility of some individuals depends upon causing harm to others, it hardly seems fair to say that their utility should be counted equally with the utility of individuals whose satisfactions are not morally offensive. Rawls also questions expensive tastes and those associated with lack of foresight or self-discipline. Thus the problem with utilities equality is not so much that utilities are difficult to observe and measure as that were we even to do so it is not clear that we could agree on what we would count as acceptable satisfactions.

But there is another important issue here. When we concern ourselves with either equality of incomes or utilities, we make equality of outcomes our focus. One might, then, avoid the problems above by focusing on equality of opportunities or that individuals are equal in having the 'same' starting points, rather than in terms of where

they end up. Rawls argues for this type of standard in connection with his first principle of justice that calls for equality in primary goods. While Sen objected to primary goods as the best measure of equality on the grounds that different individuals have different capacities for making use of primary goods, his capability approach can be seen as offering an alternative way of thinking about equality of opportunities. More accurately, capabilities equality involves both equality of opportunities and equality of outcomes, since capabilities enable individuals to do things and at the same time represent a state of achievement. This latter outcome dimension is not unrelated to utilities equality. Sen has always resisted the view that there ought to be a single set of capabilities all individuals develop. Different individuals are likely to be happier developing different capabilities, and accordingly an equality of capabilities also reflects an equality of satisfactions. But if individuals are very different in the capabilities they develop, is there any practical way of recommending an equality of capabilities?

Sen's original thinking about capabilities revolved around the idea of 'basic capability equality' (Sen 1980). When we consider extreme poverty—inadequate nutrition, housing, health care, and education—individuals differ very little in terms of their states of satisfaction (which are equally low). But to see that individuals are able to eat adequately, provide themselves housing, etc. is still to formulate capabilities equality in both equality of opportunity and equality of achievement terms. Thus the adoption of Sen's capability framework by the United Nations Development Programme in the form of the Human Development Index (HDI) makes it possible to track efforts in combating poverty as both an improvement in standards of living and as a creation of new opportunities for individuals escaping

poverty. Indeed the HDI has been used to evaluate development in the global economy by providing annual country comparisons of success in raising income, literacy, and longevity (UNDP, various years). In addition, there have been a variety of efforts to operationalize the capabilities framework in across the global economy, for example in connection with the activities of Oxfam field staffs in connection with capability development in Pakistan (e.g., Alkire 2002).

The Policy-Maker Perspective

Policy-makers, of course, may deliberate over different normative principles they wish to employ in making decisions, but at some point need to act in specific ways that reconcile possibly competing principles. Until recently, this reconciliation favored emphasis on efficiency on the grounds that the efficiency principle was clearly understood, while the equity principle was subject to competing interpretations. But closer examination of the efficiency principle has demonstrated that it includes implicit normative content that gives particular weight to equality as an interpretation of equity. The comparison of Rawls and Sen in this entry shows that there still remain significant differences in how the principle of equality is interpreted. Nonetheless, there seems to be an increasing consensus that equality not only constitutes a central policy goal, but also that it can only be achieved at the expense of efficiency is mistaken. Indeed, that efficiency includes the ideas of impartiality, freedom, and respect for individuals suggests that pursuing efficiency and equity (as equality) together is a reasonable strategy for policy-makers. That is, if pursuing efficient economic outcomes requires attention to impartiality, freedom, and respect for individuals, the way to pursue efficiency is to see that these additional normative principles function as its preconditions. From this

perspective, efficiency and equity need not be traded-off against one another as competing principles which cannot be simultaneously satisfied, but should rather be seen as compatible and complementary principles whose joint pursuit in policy requires attention to their interconnection.

Selected References

- Alkire, S. (2002) *Valuing Freedom. Sen's Capability Approach and Poverty Reduction*. Oxford University Press.
- Arrow, K. (1963 [1951]) *Social Choice and Individual Values*. Second Edition. New Haven: Yale University Press.
- Atkinson, A. (1980) "Horizontal Equity and the Distribution of the Tax Burden", in H. Aaron and M. Boskin (Editors), *The Economics of Taxation*. Washington, DC: Brookings Institution.
- Bardhan, P. (1996) "Efficiency, Equity and Poverty Alleviation: Policy Issues in Less Developed Countries", *Economic Journal*, Volume 106, pp. 1344-1356.
- Barry, B. (1965) *Political Argument*. London: Routledge and Kegan Paul.
- Bentham, J. (1970 [1789]) *An Introduction to the Principles of Morals and Legislation*. London: Athlone Press.
- Blank, Rebecca M. (2002) "Can Equity And Efficiency Complement Each Other?" *Land Economics*, Volume 9, Number 4, pp. 451-468.
- DeMartino, G. (2000) *Global Economy, Global Justice*. London: Routledge.
- Hicks, J. (1939) "The Foundations of Welfare Economics", *Economic Journal*, Volume 49, Number 4, pp. 696-712.
- Hoxby, C. (1996) "Are Efficiency and Equity in School Finance Complements or Substitutes?" *Journal of Economic Perspectives*, Volume 10, Number 4, pp. 51-72.
- Kaldor, N. (1939) "Welfare Propositions in Economics and Interpersonal Comparisons of Utility", *Economic Journal*, Volume 49, Number 3, pp. 549-552.
- LeGrand, J. (1990) "Equity Versus Efficiency: The Elusive Tradeoff", *Ethics*, Volume 100, Number 3, pp. 554-568.
- LeGrand, J. (1991) *Equity and Choice*. London: HarperCollins.
- Marshall, A. (1920 [1890]) *The Principles of Economics*. London: Macmillan.
- Musgrave, A. (1959) *The Theory of Public Finance*. New York: McGraw-Hill.
- Myles, G. (1995) *Public Economics*. Cambridge: Cambridge University Press.
- Okun, A. (1975) *Equality and Efficiency: The Big Trade-off*. Washington, DC: Brookings Institution.
- Ramsey, F. (1927) "A Contribution to the Theory of Taxation", *Economic Journal*, Volume 37, pp. 47-61.
- Rawls, J. (1971) *A Theory of Justice*. Cambridge: MA: Harvard University Press.
- Robbins, L. (1935 [1932]) *An Essay on the Nature and Significance of Economic Science*. Second Edition. London: Macmillan.
- Samuelson, P. (1986) "Theory of Optimal Taxation", *Journal of Public Economics*, 30, 137-143.
- Sen, A. (1977) "Social Choice Theory: A Re-examination", *Econometrica*, Volume 45, pp. 53-89.
- Sen, A. (1980) "Equality of What?" in S. McMurrin (Editor), *Tanner Lectures on Human Values*. Cambridge: Cambridge University Press.
- Sen, A. (1999) *Development as Freedom*. New York: Knopf.
- Sheshinski, E. (1972) "Relation Between a Social Welfare Function and the Gini Index of Income Inequality", *Journal of Economic Theory*, Volume 4, pp. 98-100.
- United Nations Development Programme. (Various Years), *Human Development Report*. Oxford: Oxford University Press.

Wagstaff, A. (1991) "QALYs and the Equity-Efficiency Trade-Off", *Journal of Health Economics*, Volume 10, Number 1, pp. 21-41.

John B. Davis
Marquette University,
Milwaukee, USA;
Universita of Amsterdam,
The Netherlands
john.davis@mu.edu
j.b.davis.uva.nl

Enterprise Net Income

Allan Young

Introduction

The proper determination of enterprise net income (or profits, as in this essay we shall use these terms interchangeably, since net income and profits are generally taken as equivalent concepts) is a problem which has perplexed accountants, economists, the business community, government entities and many others since the early stages of the industrial revolution when the importance of the undertaking became apparent to all of the constituencies relevant to the governance of the business and other affairs of enterprises. Moreover, as shall be shown, the results of the choice among the array of available and generally accepted at the time, net income determination procedures and outcomes can have significant and considerably different implications for, and a profound effect upon, an extensive number of critical governance issues germane to the interests of each of the stakeholders relevant to the business enterprise. The owners of a business need to know the profit it has earned in order to, (among many other reasons) measure the degree of the firm's success and plan for the future. Those who provide financing for enterprise activity need to be aware of the profits which have been earned in order to properly estimate the expected risks and returns from their financial venture. Workers will want an understanding of a firm's net income so as to gauge their wage and working condition demands accordingly. Governmental entities will normally assess taxation and other such enterprise levies based upon a firm's success as measured by reported net income figures. Even the community at large will look to a firm's reported net income in order to analyse the proper claims upon it based upon its being a

"good corporate citizen" as measured by the size of the positive externalities which accrue to the community from the presence of the enterprise in relation to the net income which it has been found to have earned. But throughout the decisions and positions of these various corporate stakeholders, the problem remains as to the methods and procedures by which enterprise net income is to be computed. Further, this problem is greatly exacerbated, as noted, by the differing methods and procedures utilized for the determination of enterprise net income.

Historical Perspective

We begin our analysis of net income determination with an historical overview of the development through time of the concept as perceived by accountants and economists. An understanding of the evolution of this concept is useful in attempting to grasp the meaning of the efforts of the accounting profession to come to grips with the problem of net income determination in the contemporary environment. Moreover, this historical perspective will cast light upon the governance issues noted above.

Not unexpectedly, the methods and concepts employed by the accounting profession to determine business profits (or net income) have changed over time as a function of the degree of complexity of business activity and the requirements of enterprises for external financing in order to inaugurate, carry on and expand their enterprise endeavours. Also, it is well worth noting that, as will be shown below, the area of accounting net income determination paralleled, but generally lagged considerably behind, the development of economic theory in this area.

At the dawn of the Industrial Revolution, enterprise activity was rather simple, consisting primarily of trading pursuits and cottage industry, with manufacturing

restricted to unsophisticated rudimentary forms rarely requiring substantial capital or external financing. In this enterprise environment, all that was needed for most business activities was a system of cash accounting in which particular note was taken of changes in a firm's cash position over time and an assessment made of the liquidity needs and constraints. A business climate such as this required little in the way of profit recording and reporting and, accordingly, little, if any, concern was directed in this direction, either in accounting theory or practice or by the economic theorists of the day. At that time cash transactions were the order of the day and cash flow recording represented the prime focal point of accounting concern. Moreover, even the development of economic theory was still at such an elementarily level that no real concern had yet been devoted to the determination of business profits.

Perhaps the first great and unifying economic theorist, and almost certainly the first to treat the question of the determination of business net income within a consistent framework of the economic behaviour of society was Adam Smith. Yet, although the theory of double entry bookkeeping had been rather fully developed almost three centuries before *The Wealth of Nations* was initially published in 1776, (and it is generally believed that this system of recording and classifying business transactions was in operation, if only in rudimentary form, for a far longer time previously) during this entire period no generally accepted method of computing the net income of an enterprise on a direct, conscious and consistent basis had been developed and put into practice so as to determine this net income and its fluctuations over regular time intervals. Perhaps more significantly, as practice often follows theory, no noteworthy theoretical exposition of accounting theory of net income

determination had yet been undertaken as of that time. Even in the latter part of this period prior to the publication of *The Wealth of Nations*, when the greatest progress to date was then generally made in the theoretical development and practical use of business and accounting data, the principal accounting treatises of the day typically dealt primarily with the treatment of various types of ledger accounts and the array of business transactions then common. Indeed, while some contemporary authors concerned themselves with the development of proper procedures for the temporary closing of individual accounts, others advanced accounting theory in this area a little further by considering the appropriate procedure for the formal and consistent closing of the entire ledger from time to time. But the principal concerns which preoccupy the accounting theorists of our era, basically the valuation of assets and claims upon the business and the calculation of periodic net income, took up very little space in the accounting treatises of that day and were generally touched upon only in passing in connection with the closing of the accounts. Even when, in a few rare instances, methods of stating asset values were dealt with in some slight detail, the effect of the method upon the calculation of profit was not considered.

Moreover, the practice of accounting for the determination of enterprise net income was certainly not in advance of the theoretical expositions in the area. Frequently, books were kept in such a manner that there was no attempt to separate the capital account of the proprietor or partners from the profit and loss accounts of the business. It was not uncommon to discover that not all of the assets or liabilities of a firm were included in the ledger. Hence, quite apart from the issue of the proper determination of enterprise net income at that time, its account books were frequently imperfect documents for even the

establishment of entrepreneurial valuation or net worth, either in a practical or theoretical sense. Yet in those far more simple days of business undertakings, the financial and other needs of businesses and their constituencies usually did not require the determination of profit for their effective governance. At that time business ventures were relatively small and outside equity or debt capital rarely required. And until a proper theory for the consistent determination of business net income could be developed, the practice of accounting, which has normally lagged behind theory in either case, spent little time or effort in the calculation and determination of business net income. Therefore, even in spite of the considerable technical advances in business and economic activity itself, in both in England and America, practicing accountants did not begin to attempt to determine net income upon a consistent basis until well into the nineteenth century.

However, in *The Wealth of Nations*, Adam Smith began to develop an understanding of a consistent procedure for the determination of business net income as what can be consumed or is available for consumption by an individual or economic entity without encroaching upon its capital. An accountant could then work with this definition of income for any particular business period by employing the following formulation: The accountant would determine the difference between the value of an enterprises net wealth (or net worth, as the term is more commonly used today) as of two points in time and then add back a capital consumption allowance. In essence, equipped with this basic definition of net income, accountants for the first time had a direct and consistent procedural device for its actual computation in practice. This is the so-called “Net Worth Method” of net income determination, which can be written symbolically as:

The Basic Accounting Equations

$$A_{t1} = L_{t1} + NW_{t1} \quad (1)$$

$$NW_{t1} = A_{t1} - L_{t1} \quad (2)$$

$$A_{t2} = L_{t2} + NW_{t2} \quad (3)$$

$$NW_{t2} = A_{t2} - L_{t2} \quad (4)$$

$$NW_{t2} - NW_{t1} +/\text{-} CCA_{t1-2} = NI_{t1-2} \quad (5)$$

where:

A_{t1} = Total Assets of the Enterprise as of the end of time period 1,

L_{t1} = Total Liabilities of the Enterprise as of the end of time period 1,

NW_{t1} = Total Net Worth of the Enterprise as of the end of time period 1,

A_{t2} = Total Assets of the Enterprise as of the end of time period 2,

L_{t2} = Total Liabilities of the Enterprise as of the end of time period 2,

NW_{t2} = Total Net Worth of the Enterprise as of the end of time period 2,

CCA_{t1-2} = Capital Consumption Allowance from time period 1 to time period 2,

NI_{t1-2} = Net Income of the Enterprise from time period 1 to time period 2.

Hence the translation of Adam Smith's economically-based definition of net income into accounting theory and the creation of the net worth method of income determination, as a guide to accounting practice, are marked by what has come to be known as the balance sheet approach to the determination of net income. In this method of net income computation, the balance sheet is thought of as the crystallization of the correct financial position of the enterprise and its primary financial statement from which all else of consequence relevant to the assessment of its performance and current status is derived and determined. Accordingly, under this

approach, the essential thrust of all accounting procedures is to produce a balance sheet which will properly and faithfully reflect a firm's assets, liabilities and hence its net worth (or ownership equity) as the difference between assets and liabilities as of a specific point in time. Therefore, all transactions are analysed and quantitatively assessed in terms of their effect upon the assets and liabilities of the firm, and net income, while not the primary focus of its accounting endeavours, can still be determined indirectly, through a comparison of the respective net worth's of the firm as of two points in time by the addition (subtraction) of a capital consumption allowance over these two time periods.

As an example, in 1818, F. W. Cronhlem enunciated the net worth method of net income determination in a form quite close to the one used until well into the modern era. In *Double Entry by Single* (1818), published in London, he tells us that the aim of accounting as to the determination of property values is to disclose to the owner (and presumably other enterprises stakeholders as well, as noted above) as of any point in time desired, the net value of the enterprise. The size and types of the various component elements of enterprise ownership will be in a continual state of transformation and change. However, whatever variation these elements undergo, and whether the net worth of the firm increases, diminishes, or remains stable, this value must always be equal to the sum of its parts, both positive and negative. Finally, this equality is the great essential principle of accounting.

From here, it is but a short jump to applying Smith's definition of enterprise net income to the accounting records. For once Cronhelm tells us that assets and liabilities are opposite in nature so that the net worth of the firm must always be precisely equal to the difference between them, the idea of

computing the net income of a firm during any given period of time as the difference between the net worth of these two respective periods, could not be far off. Thus, Cronhelm was able, some 42 years after the publication of *The Wealth of Nations*, to conceive of measuring the net income and losses of a business entity between two points in time as determined by the increases and decreases in its respective net worth's. Accordingly, the determination of enterprise net income as given by the difference between its net worth as of two points in time became generally known as the net worth method of net income determination.

However, perhaps reflecting the low order of importance which was given to net income figures at that time, the theoretical accounting foundation for the net worth method of computing net income was not in place and made operational until over 42 years after the publication of *The Wealth of Nations* was published and Adam Smith had initially laid down the theoretical economic basis for this method by defining net income. Moreover, there was a far greater time lag until this method of net income determination became the generally accepted accounting procedure for the computation of net income. A guide to the timing of the general acceptance of the net worth method of income determination in Great Britain is the *Companies Clauses Consolidation Act of 1845*. Sections 115-119 of this act, among other things, sought to regulate the accounting procedures of the nation's railroads. It was mandated that a balance sheet be compiled in which all assets and liabilities of the firm be computed and that the net worth of the railroad be determined as the difference between its assets and liabilities. Accordingly, given the balance sheet approach which was required to be maintained on the books and records of British railroads, it was only natural that the net income of these railroads be computed

through this financial statement as well. Eventually, almost all other enterprises of any significant size in the country began to follow the nation's railroads and accounting practice for the computation of net income began to utilize the balance sheet approach. Once this approach was adopted, the general acceptance of the net worth method of net income determination soon reached the level of general acceptance as a natural counterpart. This occurred in Great Britain in the late 1850's, more than three quarters of a century after the initial publication of *The Wealth of Nations*. General acceptance in the United States of the net worth method of net income determination came somewhat later. By mid-nineteenth century, Great Britain was far more industrially developed than the United States, and, eventually, following British practice, the net worth method of net income determination ascended to the level of general acceptance in the United States as well, as business activities became more complex and as the need for balance sheet figures in the normal conduct of business affairs became more widespread.

There were many great economic theorists in the decades immediately following Adam Smith who sought to provide a cosmological explanation for the economic affairs of mankind. But from the viewpoint of the development of accounting theory, none had as much influence in outlining and setting forth the groundwork for the next major change in the development of its precepts for the determination of enterprise net income as did Alfred Marshall. This next major change in accounting for net income determination was the development of the matching concept, wherein costs and revenues are matched and compared in order to compute accounting net income and the theoretical economic framework through which this concept was developed stemmed from the *Principles of Economics*, the first edition of

which was published by Marshall in 1890. Marshall maintained that the net income of a business should be computed by deducting from its revenue for a given period, the expenditures incurred in the production of this revenue during the same time frame. Marshall noted in his *Principles of Economics*, that in the production of revenue, businesses have to incur certain expenditures for raw materials, the hire of labour, etc., and that the true net income of the business is found by deducting these expenditures from its gross income in order to compute its net income.

Yet, although Marshall's *Principles* had gone through eight very popular editions by the late 1920's, a review of theoretical and practical accounting literature at that time suggests that the matching concept, the economic theoretical referent of which was initially developed by Marshall almost 40 years previously as his definition of net income, had not yet reached the level of general acceptance, either in accounting theory or practice. The leading accounting authorities at that time, both in Great Britain and the United States, seemed then in full agreement that the increase in net worth concept of net income determination was acceptable as the proper theoretical mode upon which to base the practice of net income determination in terms of general accounting acceptance. This was made clear in the classic, "Changing Concepts of Business Income", in Robert H. Montgomery's (1952:360ff), *Report of the Study Group on Business Income*.

For purposes of timing the point of general acceptance of the matching concept of net income determination by practicing accountants, the date September 22, 1932, is of great significance. On that day, a committee of the American Institute of Certified Public Accountants (AICPA), which was to work with the New York Stock

Exchange to adopt more effective methods of net income determination than that produced by the net worth method, sent a letter of great consequence to the Exchange. In this letter the committee rejected "the increase in net worth" concept and accepted the position that the matching concept was preferable since it emphasized the cardinal importance of net income determination as an end in itself, rather than as a derivative of balance sheet data. In other words, for the first time, net income was to be practically and separately computed through its own financial statement. Now through the matching concept, net income could be determined directly as an end in itself, by means of the newly more significant income statement. However, once again, we find a rather lengthy time gap between the derivation of a net income concept by an economist, (in this case Alfred Marshall) and the inauguration of general acceptance and application of this definition for both theoretical and practical accounting purposes.

As noted, along with general acceptance of the matching concept by accountants came a natural concomitant shift in the relative emphases given to the two major financial statements of the time, from the balance sheet to the income statement as the primary accounting reporting document. Yet with this shift in importance between the financial statements also came considerable criticism leveled at the accounting profession, both within and external to its ranks, for, among many other things, its failure to take an approach more in tune with the recognition of the increasing importance of cash flow. Cash flow represents a concept more aligned with the initial understanding of business net income in the earlier and far simpler days of enterprise activity some centuries previously. The link between cash flow and net income is clear once one realizes that the activities of the firm are focused toward the determination

of profits, which will hopefully, in time, be realized in the form of increases in cash and near cash items, such as accounts receivable and marketable securities. In a basic sense, cash flow represents the difference between the beginning and ending cash positions of a firm over two points in time. However, more sophisticated, relatively recent, understandings of the term cash flow, such as current assets (cash + marketable securities + accounts receivable + inventories + current accruals), or quick assets (cash + marketable securities + accounts receivable), or sometimes net quick assets (quick assets — current liabilities) or even working capital or net working capital (all current assets — current liabilities), have been utilized.

Various concepts of cash flow began to rise in importance during the worldwide economic depression of the 1930's, when bankruptcies, even among large-sized business entities, were common and widespread all across the globe. Cash seemed to be in great shortage, both domestically and internationally, and liquidity-based cash flow concerns were frequently of paramount importance as a means of judging the extent of enterprise illiquidity and thereby obtaining a better understanding of business status and prospects during that era of greatly depressed economic conditions.

Perhaps the most noteworthy economist critic of the accounting profession during the early 20th century was Thorstein Veblen. Of Veblen's many criticisms of accounting principles and procedures, none were more penetrating, nor perhaps more apropos, than his pointed rebuke of their reliance upon the stable monetary unit assumption and the determination of asset values, their associated accruals and the resulting net income determination procedure based upon historical cost figures. Yet it should be noted that historical cost based accounting is still largely in use at the current time and as of the

present writing, though departures from it by security analysts and others who seek to more correctly assess enterprise valuation are now commonplace highlighting the many fallacies of historical cost based accounting.

But once again, a considerable time lag is evident between the relevant writings of Veblen (*The Instinct of Workmanship*, 1914 and *Absentee Ownership and Business Enterprise in Recent Times*, 1923) and the first serious attempts by accounting theorists to face up to the obvious infirmities of net income determination techniques of the day, as exemplified by the matching concept in which costs and revenues incurred at differing points in time with perhaps widely divergent price levels were nevertheless to be compared in order to produce a resultant net income figure. Henry Sweeney's *Stabilized Accounting* in 1936 represented the first serious theoretical accounting attempt to adjust financial statements in order to take account of price level changes over time. Yet significantly, while Sweeney's work was accorded due theoretical acclaim for its comprehensiveness in attempting to deal with an extraordinarily complex, yet compelling issue, the techniques he developed have never been adopted even in part, nor have they ever even come close to reaching the status of general acceptance.

The *Study Group on Business Income* in 1952 was one of the first groups of eminent practicing accountants to really give serious attention to the problems first raised by Veblen almost forty years previously. And it wasn't until relatively recent years that the United States Securities and Exchange Commission and the accounting profession began to issue specific and concrete proposals for providing alternate reported net profit statements which took account of changing price levels.

Moreover, eventually by the 1970's, accounting authorities, in response to the

rampant inflation of the era, began requiring the reporting of fund flow figures in a statement of source and application of funds, or more commonly, a funds statement. Nevertheless, the determination of the particular fund concept upon which the statement of source and application of funds was to revolve was left to the management of the reporting firm which, as noted above, could choose as narrow a definition of funds as cash or as broad a rendering of the concept as working capital. However, by then, security analysts and others seeking to come to grips with the difficult, yet essential, problem of enterprise valuation, were virtually in uniform agreement as to the need to adjust historical cost based balance sheets and the resulting matching concept based income statements to take account of the serious infirmities of each. Additionally, and most significantly, these powerful, knowledgeable and well connected groups of large enterprise value assessors began calling for a return to cash flow concerns and techniques of enterprise valuation and then predominantly using in their work what they felt to be the more reliable financial statement: the funds statement. Accordingly, these professionals now rely far more upon the funds statement than the balance sheet or the income statement, especially when the former is derived through the use of cash as the funds concept. For they are well aware that each of the items in the balance sheet and (through accrual write offs of balance sheet accounts to the income statement and other assumptions) the entire income statement are amenable to accounting choice at the discretion of the management of the enterprise. Only cash is inviolate with respect to management choice and potential manipulation, or relatively so, and not within the control of management to pick among an array of accounting principles, each generally accepted by the profession, with frequently

widely differing reported net income determination results. So it is hardly surprising that the funds statement and cash flow concepts have risen in recent decades to ascendancy and eclipsed the balance sheet and the income statement as the primary financial document, especially for purposes of enterprise valuation.

Therefore, we have witnessed an interesting historical accounting metamorphosis of circular change in emphasis through the centuries of business history with respect to the focus of accounting thought and practice in the area of net income determination. As has been shown above, initially simple business practices required only the simplicity of cash as the prime determinant of business results. This gave way in the mid-nineteenth century to the balance sheet as the most significant financial statement as assets were felt to be the prime repository of security behind the loan capital of the day, which constituted the lion's share of enterprise financing. As an equity investing culture grew and began to take hold of business financing devices, so too did the need for separately derived net income figures and the matching concept come to the fore. However, in recent decades, the sophistication of professional investors led to a considerable disenchantment with the ease with which management can manipulate net income figures and the need for a more reliable and verifiable device for measuring enterprise success; hence the return to cash flow accounting concerns and techniques of enterprise valuation. And so historically, we have witnessed and discussed above the circular evolution from an emphasis upon cash flow accounting to the balance sheet-net worth method of income determination procedures, to the growth and decline of the matching concept as cash flow once again has risen to paramount concern as the prime measure of enterprise success and valuation.

Techniques and Practices

But beyond this interesting historical evolution of accounting thought and practice in the area of net income determination procedures, what enterprise governance considerations have both driven these accounting changes and been closely effected by them? In essence, most of these corporate governance considerations stem from an analysis of the interests of the various stakeholders or interest groups relevant to business affairs and the agency cost aspects which result when the interests of one stakeholder group is advanced at the expense of any of the others.

For the vast majority of large enterprises in the modern era, especially multinationals, the proxy system (whereby management solicits proxy votes from shareholders) effectively empowers management, which already controls the internal accounting of the firm, with the essentially unchallenged ability to select the external auditors of the firm as well. To the extent that there may well be a divergence in interests between the shareholders (or any other stakeholder group) of the firm and its management, this is much akin to allowing the home team in the case of a sporting event to choose the umpires. Further, as the degree of this divergence widens, the agency costs associated with management's choice of auditors increases as well. By management having an array of accounting principles to choose from in virtually all asset areas (save cash), the agency problem of management's choice of auditors only increases. Further exacerbating this difficulty is the frequent practice of accounting firms providing tax and management consulting services as well as auditing. The debacles of Enron and Arthur Anderson serve to underline this issue as does the felt need of the US Congress to pass the *Sarbane-Oxley Act* in 2002. Management

chooses from an array of accounting net income determination principles, all considered "generally acceptable." They then hire the auditors who may also do tax and management consulting work for the firm as well. By auditing the firm's financial statements, the auditors are, in effect, auditing the results of their own consulting recommendations. This circularity and inconsistency bespeaks of an "old boys system" rife for fraud and manipulation, and so this has occurred in a number of notorious cases in recent years with disastrous results for other affected stakeholders, and brings into clear relief serious governance issues. It also shows the significance to these corporate governance concerns of the cash flow concept of net income determination. Cash is one asset for which there is really no accounting choice and hence its reliability is even more to be valued in a world in which management's choice of auditors is virtually inviolate and very difficult, if not impossible, for other stakeholders of the firm to effectively challenge.

Now looking to the historical development of accounting net income determination outlined above in terms of governance issues, the balance sheet approach favored debt holders, especially short-term debt holders, who looked to the assets of the firm for redress of nonpayment of claims and other such grievances and rarely considered net income as a source of backing for fulfilment of their rights. Short-term debt holders do not normally have the time to await income to be realized in cash for their claims to be met. For such interests groups, the cash and then the balance sheet approaches sufficed amply. However, as the needs of businesses grew, equity financing, with attendant income statement concerns, came to the fore. Still, even for such entities, as inflation ravaged historical cost based assets, the need for a more reliable alternate became apparent.

Cash-based accounting has developed into such an alternate and is the one which results in the lowest agency costs to the various constituency groups and stakeholders of enterprises at the present time.

Cash flow based income determination is also the concept with the least opportunity for informational asymmetry between the various governance stakeholders of the firm and therefore the one with the least associated agency cost. Further, from an epistemological perspective, as the informational content in cash flow accounting is most reliable, the least adverse selection and moral hazard problems result. Finally, from a societal perspective in terms of the efficiency of the allocation of scarce capital resources, net income determination by means of cash flow concepts offers a far better alternative than either the balance sheet or the income statement approaches. Since accounting choices are avoided and the epistemological informational content of accounting data far enhanced, capital allocation decisions are much improved.

Conclusion

And so in accounting net income determination we have come full cycle, from a simple cashbased model of accounting for rudimentary business transactions applied at the dawn of the Industrial Revolution, to a more complex recording of business events as the nature of these transactions became far more complex. Eventually, the balance sheet grew to ascendancy as the embodiment of the financial affairs and prospects of the firm.

However, in the modern era, especially for purposes of net income determination, the balance sheet was eclipsed by the income statement with its matching concept. Yet, by the time the 20th century drew to a close, and the 21st began to unfold with numerous accounting related corporate debacles and a general serious and protracted equity market

downturn, analysts of securities and providers of enterprise capital, recognizing the inherent difficulties of both the balance sheet and the income statement approaches to net income determination and the attendant corporate governance concerns noted above, had developed adjustment techniques to both these financial documents which effectively returned emphasis to cash-based concepts. Accordingly, the funds flow statement, or the statement of source and application of funds, has emerged as the prime analytical tool of those seeking viable solutions to today's more complex valuation issues. As also noted, these changes over time in accounting thought with respect to net income determination have had a profound effect upon enterprise governance concerns. Finally, given that accounting is, in essence, simply a branch of economic thought, it should hardly be surprising that these changes in accounting theoretical precepts and practices have both lagged behind and were ushered in by a concomitant evolution in economic thought through the ages of business history

Selected References

- Armstrong, P. (1987) "The Rise of Accounting Controls in British Capitalist Enterprises", *Accounting, Organizations And Society*, Volume 12, Number 5, pp. 415-436.
- Arnold, P.J. (1998) "The Limits of Postmodernism in Accounting History: The Decatur Experience", *Accounting, Organizations And Society*, Volume 23, Number 7, pp. 665-684.
- Bergstrom, K.H. (1974) "Looking Back", *Management Accounting*, March, pp. 47-50.
- Bisgay, L. (1985) "MAP Statement Promulgation: A Historical Perspective", *Management Accounting*, April, 72.
- Boland, R. (1987) "Discussion of Accounting And the Construction of the Governable Person" *Accounting, Organizations and Society*, Volume 12, Number 3, pp. 267-272.
- Brief, R.P. (1965) "Nineteenth Century Accounting Error" *Journal of Accounting Research*, Spring, pp. 12-31.
- Brief, R.P. (1975) "The Accountant's Responsibility in Historical Perspective", *The Accounting Review*, April, pp. 285-297.
- Brown, R. (1968) *History of Accounting And Accountants*. Frank Cass & Co.
- Bryer, R.A. (1993) "The Late Nineteenth-Century Revolution in Financial Reporting: Accounting for the Rise of Investor or Managerial Capitalism?", *Accounting, Organizations And Society*, Volume 18, Numbers 7-8, pp. 649-690.
- Buttimer, H. (1962) "The Evolution of Stated Capital" *The Accounting Review*, October, pp. 746-752.
- Chambers, R.J. (1995) *An Accounting Thesaurus: 500 Years of Accounting*. Amsterdam: Pergamon Press.
- Cronhelm, F.W. (1978) *A New Method Of Book-Keeping. Double Entry By Single*. New York: Arno Press.
- Edwards, J.D. (1954) "The Emergence of Public Accounting in the United States, 1748-1895", *The Accounting Review*, January, pp. 52-63.
- Fleischman, R.K., and L.D. Parker. (1991) British Entrepreneurs and Pre-Industrial Revolution Evidence of Cost Management", *The Accounting Review*, April, pp. 361-375.
- Hopwood, A.G. (1987) "The Archeology of Accounting Systems", *Accounting, Organizations and Society*, Volume 12, Number 3, pp. 207-234.
- Hunt, H.G. III. and R.L. Hogler. (1993) An Institutional Analysis of Accounting Growth and Regulation in The United States" *Accounting, Organizations And*

- Society*, Volume 18, Number 4, pp. 341-360.
- Johnson, H.T. (1972) "Early Cost Accounting for Internal Management Control: Lyman Mills in the 1850s", *Business History Review*, Winter, pp. 466-474.
- Johnson, H.T. (1975) "Managerial Accounting in an Early Integrated Industry: E.I. Du Pont De Nemours Powder Company, 1903-1912", *Business History Review*, Summer, pp. 184-204.
- Johnson, H.T. (1975) The Role of Accounting History in the Study of Modern Business Enterprise", *The Accounting Review*, July, pp. 444-450.
- Johnson, O. (1981) "Some Implications of the United States Constitution for Accounting Institution Alternatives", *Journal Of Accounting Research*, Volume 19, pp. 89-119.
- Kats, P. (1930) "A Surmise Regarding the Origin of Bookkeeping by Double Entry", *The Accounting Review*, December, pp. 311-316.
- Littleton, A.C. (1931) "Early Transaction Analysis", *The Accounting Review*, September, pp. 179-183.
- Littleton, A.C. (1933) *Accounting Evolution To 1900*. New York: American Institute Publishing Co.
- Macneal, K. (1939) *Truth in Accounting*. Philadelphia: University of Pennsylvania Press.
- Marshall, A. (1920) *Principles of Economics*. 8th Ed, Macmillan and Co.
- Montgomery, R.H. (1952) *Report of The Study Group on Business Income Changing Concepts of Business Income*. New York: Macmillan and Co.
- Smith, A. (1776) *The Wealth of Nations*. London: Methuen and Co. Edited by Edwin Cannan, 1904. Fifth Edition.
- Sprague, C.E. (1972) *The Philosophy Of Accounts*. New York: Scholars Book Co. Reprint of 1908 Edition.
- Sweeney, H.W. (1936/1964) *Stabilized Accounting*. New York: Harper & Brothers.
- Vangermeersch, R. (1987) "Renewing Our Heritage: Ten Reasons Why Management Accountants Should Study the Classic Accounting Articles", *Management Accounting*, July, pp. 47-49.
- Veblen, T.B. (1914) *The Instinct of Workmanship*. New York: A.M. Kelly Publisher.
- Veblen, T.B. (1923) *Absentee Ownership and Business Enterprise in Recent Times*. New York: A.M. Kelly Publisher.
- Zeff, S.A. (1982) "Truth in Accounting: The Ordeal of Kenneth Macneal" *The Accounting Review*, July, pp. 528-553.

Allan Young
 Department of Finance,
 Syracuse University,
 New York State, USA
 aeyoung@syr.edu

Environmental Governance:

Community

Anitra Nelson

Introduction

Environmental governance refers to the socio-political aspects of making collective decisions over the use and management of natural resources. Such matters focus on environment-related political and legal rights, the regulation and responsibilities of citizens, businesses and governments, as well as social processes, values and norms, cultural ideals and models. Specifically community-based environmental governance includes various forms and levels of community participation in making and implementing decisions on the use and management of 'wild' through to constructed urban environments, including forests, river systems, dry lands, coasts, oceans, and cities.

Community-based models and techniques offer citizens power to participate in decision-making over natural resource use and management in their local area or other environments in which they are *stakeholders*. The governance aspects of community-based natural resource management (CBNRM) focus on creating mechanisms and maintaining institutions that both enable social participation and achieve sustainable environmental outcomes.

The scientific and bureaucratic approaches which characterise 'big business' and 'big government' dominated twentieth century developments in environmental management and use. In the last quarter of that century, community-based movements expanded to establish, maintain or re-establish responsibility for local resources. Much of that growth in community-based activities was due to a variety of responses to the perceived failure of transnational corporations and states to control natural resources in ways

that were both sustainable and delivered local benefits (M'Gonigle *et al* 2001). Also, economic rationalism had meant that states had devolved various responsibilities for environmental resource management either on local communities or through privatisation. Such privatisations were often controversial and citizens pressured for greater influence over such activities in their local environments.

At an international level, environmental governance principles that had been endorsed at the Rio Earth Summit in 1992 were confirmed at the World Summit on Sustainable Development in 2002. Principle 10 of the Rio Declaration on Environment and Development highlighted the importance of integrating environmental considerations in both private and public decision-making. The charter called for improved access to information and opportunities for citizens to participate in and challenge decisions made on environmental matters. In parallel—arguing the strong links between local development, grassroots democracy and poverty alleviation—international organizations such as the United Nations (UN) and a plethora of non-government organisations (NGOs) developed programs to encourage and support community activities in their local environment (Lane 1997; World Bank 1998; Durand-Lasserve and Royston 2002; Fabricius *et al* 2004; UNDP *et al* 2003). Such trends culminated in the Millenium Development Goals adopted at the UN Millenium Assembly in 2000, which identified strategies to address both poverty and environmental sustainability (UNDP *et al* 2003).

At the grassroots level, claims of Indigenous peoples to their traditional lands and broader resistance to state and/or commercial control, contributed further community-based models and programs. Issues-based environmental campaigns and

social movements, characterised by voluntary, grassroots arrangements which valued the empowerment of individuals as well as collective solidarity, flourished too (Prugh *et al* 2002). These developments demonstrated growing concerns for environmental sustainability at every spatial and governmental level. Immediate and direct local action was offered as a way to address concerns for the global environment (Mason 1999).

Permaculture (Holmgren 2002) and bioregionalism (Sale 1985) are two examples of a growing number of integrated approaches to natural resource management that stress community-based environmental awareness and local action alongside strategies of management and use based in holistic ecosystem level analyses. Bill Mollison and David Holmgren are the co-origins of permaculture, which has the objective of creating a 'cultivated ecology' for human settlements, and is a system of sustainability-based design principles and values which operate at the grassroots level and embody social justice values (Mollison 1990). In a complementary way, bioregionalism is a place-centred philosophy and practice which stresses the integration of humans within their natural landscapes and self-sufficiency and harmony within local regions (such as watersheds) of communities of plants and animals (see Planet Drum Foundation Internet site in web links below.)

Contemporary practitioners and theorists of community-based environmental governance focus on the challenges of ecological sustainability and social justice. The character and challenges of natural environments makes environmental governance distinctive from other forms of governance. Environmental rights and responsibilities are complicated by the complex character of ecosystems as well as by the diverse social, economic and political institutions which

have evolved in different regions to manage them. Also, during recent decades the increasing rate of extinctions of animal and plant species and environmental issues, such as global warming, have been analysed with reference to overpopulation of the planet and vast disparities in the material well-being of peoples within and between nations. Therefore, contemporary environmental challenges in both urban and rural areas have centred on making our everyday practices sustainable, balancing our production and consumption, and material and energy use in line with the earth's reproductive potential (Lane 1997; Prugh *et al* 2002).

While community-based organisations typically centre on neighbourhoods and regions, they are distinguished from local government in terms of style and values, as well as forms of authority. Local government is an institution of the state. Community-based governance is holistic and interdisciplinary, involving strategic public-private partnerships and communication styles, and relies less on strictly scientific approaches. Community-based governance tends to have broad aims and functions, while local government tends to be bureaucratic, hierarchical, specialist and service-oriented. In contrast to the natural resource management regimes of transnational companies and states, community-based models have tended to integrate 'soft', alternative and appropriate technologies, preferring labour-intensive and small-scale operations, encouraging local value-adding activities, and managing land and other natural resources for multiple purposes.

Global Networks

Community-based environmental governance became an increasingly significant international force during the last decades of the twentieth century. Many community-based environmental organisations have

evolved not only to promote and advocate for community-based management, but also to contribute to multinational environmental agreements that now centre on biodiversity, land, atmosphere, chemicals, hazardous wastes and the seas. Growth in electronic communication has made international networks more feasible and influential (Soeftestad 2001; UNDP *et al* 2003). International certification schemes, such as those accredited by the Forest Stewardship Council, also address issues associated with local aspirations and social equity as well as ecological integrity.

The development of global norms of good governance, local–global environmental stewardship, and social equity has been expressed in the Earth Charter and Local Agenda 21 (Mason 1999; UNDP *et al* 2003). The Earth Charter has linked local activities with global environmental impacts and participation with good environmental governance as well as social justice with ecological sustainability (Corcoran 2005). Principle 10 of Local Agenda 21 (UNCED 1992) called for informed citizens and the provision of opportunities for participation in decision-making, including access to administrative and judicial processes. Similarly, the Aarhus Convention, which came into force in 2001, recognised that a healthy environment is the right of every citizen and asserted principles of transparent and equitable decision-making, and citizens' rights to access environmental information and redress across, as well as within, national boundaries.

International developments in environmental governance have been summarised and assessed in *World Resources 2002-2004* (UNDP *et al* 2003) where principles for better governance of environmental sustainability are recommended. These principles include: adopting an ecosystem approach; committing

to the precautionary principle; improving environmental literacy; offering enabling mechanisms for broad stakeholder and citizen participation in environmental decision-making; decentralising environmental responsibilities; pursuing new partnerships; making private businesses responsible for environmental sustainability; monitoring and sanctioning environmental regulations; and strengthening international environmental cooperation. Such principles indicate strong global dimensions to CBNRM, in the form of networks, advocacy and support for activities which otherwise have a local focus (Edwards and Gaventa 2001).

Extent of CBNRM

While communal organisation in one style or another has been an historical norm the world over, the spread of capitalism and post-colonial states provided new conditions and regulations for community-based organisation. Community-based processes of environmental governance are many and varied, ranging from quasi-traditional institutions in regions such as Asia, Africa and Latin America to emergent models in white settler societies such as the USA, Canada and Australia.

Cooperatives and committees have become common forms of organising at a community level. However, community-based processes differ markedly in the extent and style of effective community power, from weak participation to substantive control. Often these differences are explained by the context within which each case operates, i.e. its specific legislative and political support, the associated material conditions and human resources available.

Despite distinct cultural, political and environmental contexts, community-based environmental developments share common political, economic and ecological challenges, such as environmental degradation, global

market forces, and inadequate government and technological support. Local Agenda 21 action plans, which focus on sustainable development, have been adopted in 113 countries by more than 6,400 local governments. However, a 2002 Gallup International poll, commissioned by the World Resources Institute, suggested that the vast majority of citizens in a range of countries wanted more information about, and opportunities to participate in, environmental decision-making (UNDP *et al* 2003.).

CBNRM embraces many different kinds of natural resources. However, CBNRM evolved most easily in ecosystems that provided humans with basic needs, such as forests (Brown *et al* 2002; Egan and Ambus 2001). Such ecosystems provided a variety of plant and animal food sources, materials for shelter, clothing, tools and artefacts for self-sufficiency and even for trade. Nevertheless, over the last century, Indigenous peoples have increasingly sought co-management with, say, mining companies and state agencies rather than pursuing exclusive rights over land. Also, greater community participation is being sought in the governance of more abstract resources such as the oceans (Costanza *et al* 1999).

Taking community forestry as an example, in 2002 around 11 per cent of the world's forests was under the control of Indigenous and non-Indigenous communities (White and Martin 2002: 7). There is a growing literature on community forestry which indicates the clear distinctions between models and outcomes in different regions as well as nations (FAO 2003).

In Africa, over 30 states had community forestry projects, involving around 5,000 communities and around 3,000,000 ha of forests (Lane 1997; Alden Wiley 2003: 17–23; Motsamai & Ntlafulang Consultants 2003). In Nepal, there are over 12,000 legal community forest user groups which manage

over 15 per cent of the country's forests and incorporate over a million households (around 20 per cent of the Nepalese population) (Pokharel 2003). India introduced a Joint Forest Management program in 1987, which was aimed at protecting forests and providing communities with economic and social benefits by devolving administrative and financial powers to legal, semi-autonomous local committees. Also, the Philippines government has instituted community forest management agreements with local communities.

Mexicans have had a long history of CBNRM. In the modern era, agrarian reforms after the Mexican Revolution (1910–17) have created 'ejidos', a tenure which allows villages usufruct rights in common and which cover around half of Mexico, including over half of the nation's forests. The Swedish forest commons and the Italian and Swiss common property and management systems also derive from models that are centuries old (Ostrom 1990). Canada's Model Forest program, which started in 1992 on large-scale forest landscapes, had multiple aims, i.e. to achieve sustainable forest management and successful private–public partnerships, to meet local-level needs and values, and to strengthen Indigenous communities (Naysmith 2003).

Programs that encourage citizen participation without involving tenure and/or have a broader, regional emphasis have evolved in popular, local and international arenas. For instance, voluntary, grass roots activities have created ecovillages, which aim to develop sustainable practices within and between neighbourhood communities (Gilman and Gilman 1991). Watershed and catchment-based restoration and stewardship, such as Landcare—which began in Australia and had included over 4,500 groups by 2000 (Carr 2002)—have developed out of government sponsored activities.

Environmental monitoring networks have been established through NGO initiatives, local action and state programs. Other developments such as the Mesoamerican Biological Corridor and the UNESCO biosphere reserves have evolved through international collaborations. Especially in these broader programs, where community power is harnessed without commensurate authority, the literature tends to criticise the extent of volunteerism and tensions between bottom-up and top-down forces. However, numerous authors have argued for a middle way with, say, state agencies setting broad standards to protect holistic ecosystems while local groups have autonomy in stewardship (Carr 2002; M'Gonigle *et al* 2001).

While there is widespread advocacy and grassroots pressure to increase community participation in urban environmental management, inadequate governance processes have been blamed for the low level of achievements (Abbott 1996). Healey (1997) has highlighted the complexities of expanding participation in urban planning and the strength of deep and enduring institutions, which tend to erode or retard the development of flexible collaborative processes. These barriers replicate those found in rural regions and developing countries where, despite efforts to be socially inclusive, sectors such as women and youth remain marginalised from and within community-based institutions (Guijt and Shah 1998; Motsamai and Ntlafalang Consultants 2003). While it is widely believed that human behaviour must alter to achieve ecological sustainability, and that this means giving all citizens opportunities to contribute in environmental decision-making, progress has been gradual to date (UNDP *et al* 2003).

Conceptual Analyses

Analysts have developed numerous frameworks and typologies for studying

community-based participatory processes. An initial reference point was the 'ladder of participation' developed by Arnstein (1969), originally to classify public involvement in social (health) services. Arnstein characterised eight kinds of power relations in distinct forms of community participation: manipulation of citizens, therapy-style relations, simple information provision, seeking views through consultation, engagement for the purpose of placation, partnerships, delegated power, and primary control by citizens.

Arnstein's model has been criticised as simplistic and unidimensional. However, it has been usefully applied to CBNRM. It has remarkable similarities with environmental typologies, such as one by Pretty (1995), which ranges through seven steps with diminishing degrees of power for community members, namely: self-mobilisation, interactivity, functional participation, participation for material incentives, consultation and manipulative participation.

As research on community-based models has grown in extent and complexity, theorists have developed various frameworks to classify the multiple forms and styles of community participation and to assess their social and environmental functions. However, most analyses still focus on the structural relations and interaction between state agencies and groups of citizens. Such frameworks tend to classify forms of participation along a scale from the de-concentration, devolution, and decentralisation of state powers to co-management and autonomous communal management (Pomeroy 1995; Motsamai & Ntlafalang Consultants 2003).

Ross *et al* (2002) have captured some of the complexity of CBNRM structures in Australia in a typology that is relevant for many countries without traditional or specific legislation for CBNRM. Their typology

includes CBM (collective tenure for multiple uses), community collective activity (voluntary monitoring and stewardship groups), organised interest groups (ENGOS, farmers' and conservationist bodies), composite stakeholder bodies (tripartite—industry, government and community—catchment management committees), shared/co-management (joint, e.g. Indigenous—state, management of national parks), and stakeholder-based planning/ negotiations (specific purpose short-term collaborations, e.g. citizens' juries).

Today, community-based models are characterised by flexible, experimental, and complex sets of relationships. For this reason it is simplistic and misleading to characterise the central dynamics as necessarily oppositional, i.e. between private—community tenures and/or state—community control (Richards 1997; Ross *et al* 2002). Private companies and landholders as well as international organisations, multilateral agencies and NGOs, have become significant actors and influences in CBNRM. Many communities, including Indigenous ones, have developed partnerships with private companies or have directly managed commercial cooperatives with special characteristics (Mayers and Vermeulen 2002; Scherr *et al* 2002).

Theorising on contemporary developments, M'Gonigle *et al* (2001) have proposed a holistic, legal and integrated 'community ecosystem trust' (CET) model with strong and transparent sustainability values related to environmental and social well-being. The CET was developed as an ideal model, commissioned by the Canadian government to propose legislative reforms for instituting healthy and enduring CBNRM in Canada. However, it drew on practical, real and successful models from all over the world. Significantly, the CET model features best

practice and performance-based regulation for ecological and social outcomes (M'Gonigle *et al* 2001).

The citizen science team of the Australian Coastal Cooperative Research Centre presents online a set of principles, a diverse toolbox of techniques and an annotated bibliography on establishing, enhancing and maintaining participatory processes. Other materials, and a gateway to Internet resources, is offered online as part of the Calabash project, managed by the South African Institute for Environmental Assessment (SAIEA), funded by the World Bank and the Canadian International Development Agency (see web links below).

'Community' and 'Environment'

Most discussions of community-based environmental governance centre on questions associated with concepts of community and governance practices. Cynics, in particular, stress that transparency and accountability are as crucial for community-based as other forms of environmental governance and that evaluations of environmental outcomes must be rigorous.

Debates and theories centre on the principles and aims of CBNRM, such as the appropriate balance between issues of social equity and environmental justice, especially when social needs conflict with broader ecosystem health. A familiar example is the collection of firewood in regions where it is the cheapest source of fuel for the poorest households but causes environmental damage. Consequently, the role of CBNRM in overcoming poverty is questioned (Fabricius *et al* 2004). Some of the deepest social conflicts have evolved over conflicting environmental values and there is a growing literature on conflict management techniques as well as conflict resolution (O'Leary and Bingham 2003).

Analysts are concerned with the constitution and legitimacy of distinctive political structures of CBNRM and the latent as well as effective power of community members. Such concerns tend to focus on whether processes are effective and democratic. Especially in regions where CBNRM is supported by the state, there have been numerous questions raised over whether representatives account for all community members' interests, including women and youth. Other studies centre on whether community structures and powers are sufficient and efficient in terms of engaging and negotiating with external (state and private sector) forces.

Theoretically, the diverse, inclusive, and collaborative character of community-based models makes them complementary to the holistic and integrated nature of ecosystems and specifically suited to appropriate, adaptive management according to local ecological conditions and social needs. However, community-based models have been criticised as insular, wilful, self-interested, ineffective, undemocratic, and exploitative of nature (Kellert *et al* 2000).

Clearly, community-based models do not in, of and by themselves ensure environmental (or social) sustainability. Cultural, political and economic contexts are critical in determining social and ecological efficiency. Gibson *et al* (2003) stress the importance of monitoring and sanctions for social rules. M'Gonigle (1998) emphasises the needs and integrity of both human communities and their environs, implying the relative autonomy of local governance.

The complexity and scale of the specific ecosystem in question are determining factors in selecting an appropriate governance model. The subsidiarity principle is often invoked. This principle refers to governing a resource at an appropriate, frequently local, level or dividing responsibilities for distinct aspects of

environmental management between parties, such as local communities and national agencies.

'Nature' is a cultural concept. Perceptions of, and values attributed to, the 'environment' mould governance structures and community-based visions, models and processes for environmental management and use. Many Indigenous communities living on traditional lands are involved in CBNRM. Indigenous communities have been characterised as more sympathetic to ecosystem and human needs than scientific-bureaucratic and capitalist forms of management (Knudtson and Suzuki 1992). The trend in conservation to exclude people from protected area national parks in the late twentieth century has given way to more inclusive local management and use (Ghimire and Pimbert 1997). A dialogue between Indigenous and western scientific perspectives has evolved (Michel and Gayton 2002). However, Indigenous peoples face similar issues as non-Indigenous communities in negotiating secure communal tenure, a beneficial relationship with state and industry, and enhancing ecological sustainability while meeting their members' basic needs (Lane 1997; Durand-Lasserve and Royston 2002).

Models and Methods

Community-based participation relies on enabling mechanisms and structures for co-ordinating collective discussions, decision-making and activities. Key issues include: the adequate collection and sharing of information; joint analyses of environmental opportunities, challenges and community members' needs and desires; access to training in key skills and specialist advice; the provision of resources, including materials and technology; efficient negotiating processes for engaging state authorities and other powerful stakeholders (Hemmati 2002); effective processes for consensual decision-making and conflict resolution; delegating

responsibilities and sharing the practical tasks of implementation; and continuous monitoring and evaluation to inform future plans.

Methods developed to support and structure CBNRM include: participatory rural appraisal and rapid rural appraisal (Chambers 2002; Jackson and Ingles 1998); livelihood analysis and baseline surveys (World Bank 1996); participatory action research and learning (Rahman 1993; Chambers 2002); appreciative inquiry and appreciative participatory planning and action (New Paradigm Consulting 2003); and environmental impact assessment (Motsamai and Ntlafulang Consultants 2003). Action learning and action research are well-acknowledged educational methods for application in sustainable development (Tilbury and Wortman 2004).

The diversity of stakeholders and ideal of involving all in discussions and decision making has raised many problems for governance and related processes of conflict resolution (UNDP *et al* 2003). Aslin and Brown (2002) have produced a comprehensive kit of tools and techniques for community engagement in natural resource management with an excellent annotated bibliography and list of relevant Internet sites.

While community-based methods favour low-technology and labour-intensive local development, new technologies are employed too. Geospatial information technology (geographic information systems, remote sensing and visualisation tools for scenario development and forward planning) offers opportunities for communities to profile their environment and social characteristics, create scenarios and monitor progress. Charettes involving public participatory planning support systems (PPSS) are increasingly employed in the USA and the UK (see PlaceMatters Internet site in web links below).

States grant and protect property and common pool resources, such as use-rights to water, minerals and fish. Devolution of centralised state powers has occurred mainly through: 1) privatisation, which generally curbs or eliminates community participation; 2) decentralisation of power to local government; 3) collaborative or co-management between government and community groups; 4) the devolution of control to local user groups (the strict form of CBNRM). The quality or extent of participation is most often judged by the powers that are given or derived in the process (Nelson and Wright 1995). All forms can include private partners and subjection to market forces. Authors such as Koontz *et al* (2004) and Lahri-Dutt (2004) have pointed out that co-management and even CBNRM tend to be dominated by government policies and influential bureaucrats or extension officers who favour scientific approaches and are held accountable for collaborative ventures.

Even though most models directly support capitalism and private property, many perceive community-based environmental management to be an 'alternative', non-capitalist or even anti-capitalist, form. The dominance of market forces has challenged CBNRM ideologically and structurally. Many debates over CBNRM have been cast in highly politicised and oppositional capitalist and anti-capitalist paradigms. In practice, private property tended to create its opposite, i.e. public or 'no-one's' property, terra nullius and wilderness. Capitalist notions of private property rights and responsibilities conflict with notions of common property (such as commons, common pool resources, public land, and state property) as well as the criteria which members of collectives tend to value. For instance, Hardin's (1968) narrative of overexploitation of natural resources, the 'tragedy of the commons', assumed self-

interested individuals acting without regard for the environment, and evoked strong criticism (Cox 1985; Ostrom 1990).

Many forms of community-based environmental governance need to balance conflicting responsibilities and claims associated with effective environmental stewardship. This includes the qualitative and quantitative aspects of deriving social and economic benefits from local resources as well as acknowledging various levels or scales of social authority depending on the resource in question. Continued development and experimentation with CBNRM is likely to result in more diverse and hybrid (common/private) forms of management and a greater exposure to and understanding of the more subtle aspects of communal, cooperative and joint management.

Conclusion:

Global and Future Significance

Discussions of governance in CBNRM tend to centre on how passive or active community members are, i.e. the terms, processes and conditions of individual engagement internally and the collective engagement with external forces. These distinctions are more apparent to the extent that undemocratic or low-level (i.e. elitist, 'representative' or 'formal') democracy persists. In contrast, participatory processes and active citizens are implicit in and essentially define 'substantive democracy'. Exactly how the interests and energies of community members can be harnessed to constructively contribute to collective social goals and ecological sustainability is a moot point. There are numerous critics of community-based environmental governance. However, there is general agreement that progress in substantive democracy and social justice, and in making environmental practices more sustainable, relies on greater participation by community members in the management and

use of environments that they impact on in daily productive and consumptive activities.

The sustainability literature heralds a revolution, or at least a planetary transition, that implies greater participatory processes in the future (Raskin et al 2002). Directions for further innovation are offered by approaches such as permaculture, proposals such as the community ecosystem trust, and the Earth Charter and eco-spiritual movements. Similarly, new information and communication technologies focussing on the Internet and based in geospatial visualisation tools have the potential to connect and inform all kinds of communities across the globe. Therefore, it is possible that in the future the concept 'community' will be less associated with a local focus and incorporate a 'global citizen' and 'steward of the planet' focus on local-global governance concerns and processes (Edwards and Gaventa 2001; Jasanoff and Long Martello 2004). In this way some of the social and environmental contradictions of community-based environmental governance might well be constructively worked through and overcome.

Selected References

- Abbott, J. (1996) *Sharing the City: Community Participation in Urban Management*. London: Earthscan.
- Alden Wiley, L. (2003) "From Meeting Needs to Honouring Rights: The Evolution of Community Forestry", in FAO, *XII World Forestry Congress Quebec City (Canada), September 2003. Proceedings A—Forests for People*: 17–23.
- Arnstein, S.R. (1969) "A Ladder of Citizen Participation", *Journal of the American Institute of Planners*, 35, 216–24.
- Aslin, H.J. and V.A. Brown (2002) *Terms of Engagement: A Toolkit for Community Engagement for the Murray-Darling Basin*. Canberra: Bureau of Rural Sciences.

- Brown, D.; Y. Malla, K. Sckrekenberg and O. Springate-Baginski. (2002) *From Supervising 'Subjects' to Supporting 'Citizens': Recent Developments in Community Forestry in Asia and Africa*, Natural Resource Perspectives 75. London: Overseas Development Institute.
- Carr, A. (2002) *Grass Roots and Green Tape: Principles and Practices of Environmental Stewardship*. Sydney: Federation Press.
- Chambers, R. (2002) *Participatory Workshops: a Sourcebook of 21 Sets of Ideas and Activities*. London: Earthscan.
- Corcoran P.B. (2005) (Editor) *Toward a Sustainable World: the Earth Charter in Action*. Amsterdam: KIT Publishers.
- Costanza, R.; F. Andrade, P. Antunes, M. van den Belt, D. Boesch, D. Boersma, F. Catarino, S. Hanna, K. Limburg and B. Low. (1999) "Ecological Economics and Sustainable Governance of the Oceans", *Ecological Economics*, 31, 171–87.
- Cox, S.J. (1985) "No Tragedy on the Commons", *Environmental Ethics*, 7, 49–61.
- Durand-Lasserve, A. and L. Royston (2002) *Holding their Ground: Secure Land Tenure for the Urban Poor in Developing Countries*. London: Earthscan.
- Edwards, M. and J. Gaventa. (2001) (Editors) *Global Citizen Action*. London: Earthscan.
- Egan, B. and L. Ambus (2001) *When There's a Way, There's a Will. Report 2: Models of Community-Based Natural Resource Management*. University of Victoria, BC: Eco-Research Chair of Environmental Law and Policy.
- Fabricius, C.; E. Koch, S. Turner and H. Magome. (2004) (Editors) *Rights, Resources and Rural Development: Community-Based Natural Resource Management in Southern Africa*. London: Earthscan.
- FAO. (2003) *XII World Forestry Congress Quebec City (Canada) September 2003 Proceedings A—Forests for People*. New York: Food and Agriculture Organisation, Forestry Department.
- Ghimire, K.B. and M.P. Pimbert (1997) (Editors) *Social Change and Conservation: Environmental Politics and Impacts of National Parks and Protected Areas*. London: Earthscan.
- Gibson, C.; J. Williams and E. Ostrom. (2003) "Rule Enforcement and Local-Level Forest Management", in FAO (2003) *XII World Forestry Congress Quebec City (Canada) September 2003 Proceedings A—Forests for People*: 165–75.
- Gilman, D. and R. Gilman. (1991) *Eco-Villages and Sustainable Communities*. Bainbridge Island, US: Context Institute.
- Guijt, I. and M.K. Shah (1998) *The Myth of Community: Gender Issues in Participatory Development*. London: Intermediate Technology.
- Hardin, G. (1968) "The Tragedy of the Commons" *Science*, 162, 1243–48.
- Healey, P. (1997) *Collaborative Planning: Shaping Places in Fragmented Societies*. Houndmills/London: Macmillan Press.
- Hemmati, M. (2002) *Multi-Stakeholder Processes for Governance and Sustainability: Beyond Deadlock and Conflict*. London: Earthscan and UNED Forum.
- Holmgren, D. (2002) *Permaculture: Principles and Pathways Beyond Sustainability*. Hopburn, VIC: Holmgren Design Services.
- Jackson, W.J. and A.W. Ingles. (1998) *Participatory Techniques for Community Forestry: A Field Manual*. Gland, CH: AusAID, IUCN and WWF.
- Jasanoff, S. and M. Long Martello. (2004) (Editors) *Earthly Politics: Local and Global in Environmental Governance*. Cambridge, MA: MIT Press.
- Kellert, S.R.; J.N. Mehta; S.A. Ebbin and L.L. Lichtenfeld. (2000) "Community Natural

- Resource Management: Promise, Rhetoric and Reality”, *Society and Natural Resources*, 13, 8, 705–15.
- Knudtson, P. and D. Suzuki (1992) *Wisdom of the Elders*. Toronto: Stoddart.
- Koontz, T.M.; T.A. Steelman, J. Carmin, K.S. Korfmacher, C. Moseley and C.W. Thomas. (2004) *Collaborative Environmental Management: What Roles for Government?* Washington DC: Resources for the Future Press.
- Lahiri-Dutt, K. (2004) “I Plan, You Participate”: A Southern View of Community Participation in Urban Australia’, *Community Development Journal*, 39, 1, 13–27.
- Lane, C. (1997) (Editor) *Custodians of the Commons: Pastoral Land Tenure in Africa*. London: Earthscan.
- Mason, M. (1999) *Environmental Democracy: A Contextual Approach*. London: Earthscan.
- Mayers, J. and S. Vermeulen. (2002) *Company-Community Forestry Partnerships: From Raw Deals to Mutual Gains?* London: International Institute for Environment and Development (IIED).
- Michel, H. and D.V. Gayton. (2002) (Editors) “Linking Indigenous Peoples’ Knowledge and Western Science” in *Natural Resource Management Conference Proceedings*. Kallops, BC: Southern Interior Forest Extension and Research Partnership.
- M’Gonigle, M. (1998) “Living Communities in a Living Forest: Towards an Ecosystem-Based Structure of Local Tenure and Management”, in C. Tollefson (Editor), *The Wealth of Forests: Markets, Regulation and Sustainable Forestry*. Vancouver: UBC Press, 152–85.
- M’Gonigle, M.; B. Egan and L. Ambus (2001) *When There’s a Way, There’s a Will. Report 1: Developing Sustainability Through the Community Ecosystem Trust*. University of Victoria, BC: Eco-Research Chair of Environmental Law and Policy.
- Mollison, B. (1990) *Permaculture: A Practical Guide for a Sustainable Future*. Washington DC: Island Press.
- Motsamai, B. and Ntlafulang Consultants. (2003) *Situation Assessment of Participation of Civil Society in Environmental Assessment in South Africa*. Windhoek: Southern African Institute for Environmental Assessment.
- Naysmith, J.K. (2003) “Canada’s Model Forest Program: Building on Success”, in FAO (2003) *XII World Forestry Congress Quebec City (Canada) September 2003, Proceedings C — People and Forests in Harmony*, 101–07. New York: Food and Agriculture Organisation, Forestry Department.
- Nelson, N. and S. Wright. (1995) *Power and Participatory Development: Theory and Practice*. London: Intermediate Technology.
- New Paradigm Consulting. (2003) *Appreciative Inquiry*. www.new-paradigm.co.uk/Appreciative.htm
- O’Leary, R. and Bingham L.B. (2003) *The Promise and Performance of Environmental Conflict Resolution*. Washington DC: Resources for the Future Press.
- Ostrom, E. (1990) *Governing the Commons: The Evolution of Institutions for Collective Action*. New York: Cambridge University Press, Cambridge.
- Pokharel, B. (2003) “Contribution of Community Forestry to People’s Livelihoods and Forest Sustainability: Experience from Nepal”, in FAO (2003) *XII World Forestry Congress Quebec City (Canada) September 2003 Proceedings A — Forests for People*. New York: Food and Agriculture Organisation, Forestry Department.
- Pomeroy, R.S. (1995) “Community-Based and Co-Management Institutions for Sustainable Coastal Fisheries Management

- in Southeast Asia”, *Ocean and Coastal Management*, 27, 3, 143–62.
- Pretty, J.N. (1995) “Participatory learning for sustainable agriculture”, *World Development*, 23, 8, 1247–63.
- Prugh, T.; R. Costanza and H. Daly (2002) *The Politics of Global Sustainability*. Washington DC: Island Press.
- Rahman, A. (1993) *People’s Self-Development: Perspectives on Participatory Action Research*. London: Zed Books.
- Raskin, P.; T. Banuri, G. Gallopín, P. Gutman, A. Hammond, R. Kates and R. Swart. (2002) *Great Transition: The Promise and Lure of the Times Ahead*. Boston, MA: Stockholm Environment Institute.
- Richards, M. (1997) “Common Property Resource Institutions and Forest Management in Latin America”, *Development and Change*, 28, 1, 95–117.
- Ross, H., M. Buchy, W. Proctor (2002) ‘Laying Down the Ladder: A Typology of Public Participation in Australian Natural Resource Management’, *Australian Journal of Environmental Management*, 9, 4, 205–17.
- Sale, K. (1985) *Dwellers in the Land: A Bioregional Vision*. San Francisco: Sierra Club Books.
- Scherr, S.J.; A. White and D. Kaimowitz. (2002) *Making Markets Work for Forest Communities*. Washington DC: Forest Trends.
- Soeftestad, L. (2001) *Community-Based Natural Resource Management: Knowledge Management and Knowledge Sharing in the Age of Globalization*. Kristiansand, Norway: CBNRM. www.cbnrm.net
- Tilbury, D. and D. Wortman. (2004) *Engaging People in Sustainability*. Cambridge, UK: Commission on Education and Communication, IUCN, Gland.
- UNCED. (1992) *Agenda 21*, United Nations Conference on Environment and Development. Geneva: UNEP.
- White, A. and A. Martin (2002) *Who Owns the World’s Forests?* Washington DC: Forest Trends.
- World Bank. (1996) *The World Bank Participation Sourcebook*. Washington DC: World Bank.
- World Bank. (1998) *The International Workshop on Community-Based Natural Resource Management. Workshop Report*. Washington, DC.: World Bank.
- UNDP, UNEP, WB and WRI. (2003) *World Resources: 2002–2004: Decisions for the Earth: Balance, Voice and Power*. Washington DC: World Resources Institute.
- Websites**
- Aarhus Convention Clearing House. <http://aarhusclearinghouse.unece.org>
- Calabash Project. (South African Institute for Environmental Assessment) www.saiea.com/
- Centre for Biodiversity and Indigenous Knowledge. www.cbik.ac.cn
- Citizen Science. (Coastal Cooperative Research Centre) www.coastalzone.crc.org.au
- Community Based Collaboratives Research Consortium. (CBCRC). www.cbrc.org/
- Community-Based Natural Resource Management Network. www.cbnrm.net/index.html
- Community Stewardship Network. (Sonoran Institute) www.sonoran.org
- Earth Charter. www.earthcharter.org
- Global Caucus on Community-Based Forest Management. www.gccbfm.org/.
- Global Ecovillage Network. gen.ecovillage.org
- IDS Participation Resource Centre. www.ids.ac.uk/ids/particip/
- International Association for Public Participation. (IAP2) www.iap2.org

International Institute of Sustainable Development. (IISD) www.iisd.org/ai/
International Network of Forests and Communities. (INFC)
www.forestsandcommunities.org
PlaceMatters.com. www.placematters.com
Planet Drum Foundation.
www.planetdrum.org
Polis Project on Ecological Governance.
www.polisproject.org
Resource Centres for Participatory Learning and Action. www.rcpla.org
United Nations Environment Program. (UNEP) Register of Multinational Environmental Agreements.
www.unep.org
UNESCO's Programme on Man and the Biosphere. (MAB) www.unesco.org/mab
World Bank. Participation and Civic Engagement. www.worldbank.org/participation
World Conservation Union. (IUCN)
www.iucn.org
World Resources Institute. Access Initiative.
www.wri.org/governance

*Anitra Nelson
Australian Housing & Urban Research
Institute, RMIT University
Melbourne, Australia
anitr@aapt.net.au*

Environmental Justice and Equity

Jouni Paavola

Introduction

Environmental justice was brought to wider attention by the environmental justice movement, which emerged from local environmental conflicts over the pollution of air, water and soil; hazardous and toxic wastes; and the siting of locally unwanted land uses and facilities in the United States in the 1980s (see Bryant and Mohai 1992; Bullard 1990, 1999). Environmental justice movement demonstrated that racial and ethnic minorities were exposed to disproportionate environmental hazards and burdens in the United States. Many other developed and developing countries have since witnessed the emergence of their own environmental justice movements (McDonald 2002; Shiva 1999).

Environmental justice has also become a subject of keen scholarship in philosophy, political science, economics, geography and sociology in the last few years (see Attfield 1999; Dobson 1998, 1999; Gleeson and Low 2001; Low and Gleeson 1998; Schlosberg 1999; Shrader-Frechette 2002). In the past, environmental research in these disciplines was often preoccupied either with sustainable development – fair allocation of resources between the present and future generations – or the obligations of humans to non-humans. The concerns of environmental justice movement have reminded that the environment is often a site of social injustice within each generation. In what follows, the emergence of environmental justice movement and the recognition of environmental injustices are discussed first. Conceptual dimensions of distributive and procedural environmental justice will then be reviewed and used for examining policy developments relevant to environmental justice. The conclusions discuss governance

implications of environmental justice concerns.

Environmental Justice Activism

Environmental justice activism and empirical research on environmental justice demonstrated in the 1980s and 1990s that environmental amenities, burdens and hazards are often highly unequally distributed. Yet environmental injustices and activism have longer roots. The poor and minorities suffered higher mortality and morbidity rates because of deficient sanitation and other environmental hazards in the 19th century (Szreter and Mooney 1998). Local environmental conflicts were sparked by slaughterhouses, rendering establishments, gas works and malodorous millponds because they were considered a health threat in the 19th century (Paavola 2004). Later many local communities organized themselves to oppose industrial water use and air and water pollution (Steinberg 1991; Stradling 1999). While this early activism often had white middle-class origins, conflicts also emerged in the workplace over environmental hazards such as radium (Neuzil and Kovarik 1996).

The origins of the contemporary environmental justice movement are often traced to the 1982 protests over the dumping of polycarbonated biphenyls (PCBs) in a predominantly minority community in Warren County, North Carolina (Bryant and Mohai 1992:2). The US General Accounting Office (1983) responded to the controversy by conducting a survey which demonstrated that in the Southern states the majority of hazardous waste landfills were located in minority communities. A few years later the Commission for Racial Justice of the United Church of Christ (1987) found that the siting of hazardous waste facilities is closely associated with race in the whole country. The US Environmental Protection Agency admitted in 1992 that low-income and

minority groups are disproportionately exposed to lead, air pollutants, hazardous waste facilities, contaminated fish and agricultural pesticides in the workplace (EPA 1992). Environmental justice received even stronger recognition when President Clinton issued Executive Order 12898 on federal actions to address environmental justice in minority and low income populations in 1994.

Empirical research has substantiated the disproportionate exposure of minority and low-income populations to environmental hazards. In 1992, a third of Hispanic Americans lived in areas where the National Ambient Air Quality Standards (NAAQS) were not attained for particulate emissions. Less than 15 percent of white population lived in such areas. While 6 percent of white population was exposed to lead concentrations that exceeded the NAAQS in 1992, 9 percent of African Americans and 18 percent of Hispanic Americans were exposed to such lead concentrations (Institute of Medicine 1999:15). The average African American lives in a county with 60 percent higher emissions of toxic air pollutants than the county in which average white American lives (Perlin, Setzer et al 1995). The average Hispanic American lives in a county with over 100 percent higher toxic air pollutant emissions than the county of residence of an average white American. Environmental hazards translate to adverse health outcomes among the non-whites. New York's Hispanics are three times as likely to be hospitalized and to die because of asthma than whites (Institute of Medicine 1999:21). In California, non-white population has up to 50 percent higher life time cancer risk than the white population because of exposure to higher concentrations of air pollutants (Morello-Frosch, Pastor et al 2002).

While empirical research on environmental justice has its longest roots in North America,

similar results are being obtained elsewhere. In the UK, poorer communities often host large industrial facilities and suffer disproportionately from emissions of known carcinogens (Stephens, Bullock et al 2001). Ethnic minorities are exposed to significantly higher concentrations of carbon monoxide and nitrogen dioxide than the white population (Brainard, Jones et al 2002:709) and low income communities in general are burdened by these emissions which originate from transport (Mitchell & Dorling 2003). Low income communities have typically low car ownership and yet, in addition to being burdened by pollution, they also suffer disproportionately from traffic deaths, especially those of children (Stephens, Bullock et al 2001). Elderly and other disadvantaged lack access to technologies, fuels and social networks which are important for coping with environmental stress. They suffer excess deaths because of exposure to cold during the winter (Stephens, Bullock et al 2001). Coping with heat also requires unequally distributed coping assets. The 1995 Chicago heat wave resulted in over 700 excess deaths, many of whom were old black males living alone (Klinenberg 2002).

Benefits of resource use and the burdens of adverse environmental impacts are also highly unequally distributed in many developing countries and this inequality often has an ethnic dimension to it. For example, in South Africa, apartheid policies included the taking of indigenous rights to land and other natural resources, siting of locally unwanted and hazardous facilities to non-white neighborhoods, and the denial of access to environmental resources and amenities (McDonald 2002). The Government of Papua New Guinea benefited from the giant Panguna copper mine in Bougainville while leaving the Bougainvilleans without compensation for appropriated land and environmental damages—an injustice which

resulted in civil war in 1993 (Denoon 2000). In Kenya and Tanzania, colonial rulers and post-independence governments favored sedentary agriculturalists over the Maasai pastoralists, creating a tension which still today erupts in deadly conflicts when droughts drive pastoralists to lowlands.

The patterns of inequality are the same in the international context. Climate change impacts will burden developing countries disproportionately, although they have not contributed to the problem (Adger et al. 2005; Anand 2004). Developing countries (particularly local communities) also bear disproportionate costs for protecting biodiversity (Neumann 1998; Brockington 2002). Moreover, developing countries produce primary and labor intensive products for export, which has detrimental effects on occupational and environmental health and environmental quality. For example, pesticides that cannot be marketed or used in the developed countries are still being manufactured and exported to developing countries (Perfecto 1992; Shrader-Frechette 2002). They are also often the destination of hazardous wastes (Anand 2004; Mpanya 1992).

To summarize, activism and research have brought up important observations regarding environmental injustices. They have already resulted in policy changes in the United States and the European Union which will be discussed in greater detail below. However, first a closer look at conceptual issues is needed to integrate environmental justice concerns to contemporary environmental governance.

Varieties of Environmental Justice

Just inter-generational allocation of resources and sustainable development are established themes in philosophy and economics. Many economists readily admit that we have a duty to invest a share of the proceeds from the use

of environmental resources for augmenting the stocks of man-made capital, in order to maintain the possibilities of future generations to enjoy equal or higher level of welfare than is currently enjoyed. They argue however that the discounting of future costs and benefits is necessary for intergenerational justice: future generations are argued to be wealthier than we are because they will be able to take advantage of productivity-enhancing investments and innovations (Beckerman & Pasek 2001; Little 2002). Economists typically deny any more specific rights and duties regarding the future generations, as we purportedly do not know what their preferences will be.

The critics have pointed out that the transformation of natural capital into man-made capital is often irreversible and that the two are not completely substitutable: some minimum amount of natural capital is needed for the sheer physical existence of future generations. The critics have suggested several principles that modify the basic rule of leaving the future generations at least as well off as we are. One approach suggests the use of *safe minimum standards* to avoid irreversible losses (Ciriacy-Wantrup 1968; Farmer and Randall 1998). Others have suggested setting off a minimum amount of critical natural capital (Ekins et al. 2003). The third alternative is the *precautionary principle*. The Rio Declaration of the United Nations Conference on Environment and Development (UNCED) defined it as requiring that “[w]here there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation”. Others, such as Norton (2002) and Sagoff (1988), deny that the lack of information about the preferences of future generations relieves us from a duty to protect the environment today. They argue that we have

a paternalistic responsibility to conserve nature today in order to cultivate environmental preferences in future generations. Another line of argumentation extends the duties of present and future generations to non-humans (Attfield 1999).

Environmental justice activism has drawn attention to intra-generational justice in environmental matters. Indeed, a comprehensive approach to environmental justice must include both intragenerational and intergenerational issues. Moreover, environmental justice should be acknowledged to have both distributive and procedural dimensions, the former relating to the incidence of environmental benefits, burdens and hazards and the latter relating to the way in which environmental plans and decisions are made. Environmental justice does not encompass the relationships between the humanity and the non-human world. Low and Gleeson (1998) have proposed that these relationships are the domain of *ecological justice*.

There are several approaches to environmental justice. For instance, theories of justice can be characterized as cosmopolitan and communitarian ones, the former arguing that justice is not dependent on time, space and culture and the latter making the case for the opposite (Attfield 1999). There are also utilitarian, other (non-utilitarian) consequentialist and deontological approaches environmental justice. In what follows, these three approaches will be discussed in turn.

The utilitarian approach and its variants such as those based on social welfare functions associate justice with the maximization of an aggregate good such as utility or social welfare (Kolm 1996; Little 2002): unequal distribution of environmental benefits and burdens is not necessarily unjust, provided that deviations from equality increase utility. Correspondingly,

redistribution is warranted when it increases utility or social welfare. The view of this approach that changes in utility can adequately represent all distributive impacts has uncomfortable ramifications. Economic benefits justify sacrifices in health or safety if the former make, on the balance, a larger contribution to utility or social welfare than the latter. Moreover, if money is used for commensurating impacts, the interests of involved parties are weighed by their ability to pay. This reasoning informed the infamous 1992 Lawrence Summers memo, according to which the World Bank should promote the flight of dirty industries to developing countries. Summers reasoned that adverse environmental and health impacts would cause only modest losses in developing countries because of small foregone earnings, because the costs of increasing pollution levels would initially be low, and because there would be little demand for environmental amenities as a result of low levels of income. The same logic underlies an assessment according to which the value of a Russian's life is five percent of that of an American (Larson et al. 1999:1815). Despite these problematic features, utilitarian arguments have informed important environmental reforms. For example, public health campaigners used them to promote safe water in the early 20th century. Investments in safe water supply reduced mortality due to water-borne diseases by 99 percent in a few decades and also improved social welfare (Meeker 1974; Paavola 2004).

There are many other consequentialist approaches to justice in addition to utilitarianism and social welfarism. Non-utilitarian consequentialist approaches share an emphasis on equality as the hallmark of justice but different views exist regarding how equality ought to be defined. For example, social justice could be defined in terms of equality of resources, income,

consumption or “satisfaction”. Unlike utilitarianism and its variants, non-utilitarian consequentialist approaches usually recognize the existence of multiple goods and that equality thus has to be multidimensional (Kolm 1996). For example, Rawls (1971) defined the original position as one characterized by equality in terms of basic goods such as income, power, status and self-esteem. He then proceeded to argue that self-interested, risk-averse agents would support the maximin principle under the “veil of ignorance” about their actual position in the society. Walzer’s (1983) notion of complex equality in turn seeks to address the challenge of incommensurable areas of equality by suggesting that a minimum requirement of justice is the absence of domination by one group across spheres of justice.

In environmental justice, equality could serve as a standard for just access to environmental resources and just exposure to environmental burdens – this seems to be the view of many environmental justice scholars. However, deviations from narrowly defined equality can be justified by responsibility (or its absence), need, capacity or desert (Barry 1998; Shrader-Frechette 2002). Qualification of equality renders justice capable of reflecting the particularities of issues and contexts. This is especially important when some distributive impacts – say health effects and risks – are arguably more important or fundamental than others, such as economic benefits (Miller 1998). Moreover, simple equality does not go far enough for true environmental justice in many instances. For example, a vulnerable group which is not responsible for or contributing to its environmental burdens should be entitled to stronger protection of its interests than what would be required under mere equality. This is the situation with the Least Developed Countries with regard to climate change, for example (Adger et al. 2005). Maximin and

leximin (lexicographic preferring of subsequently most disadvantaged groups; see Kolm 1996) principles would call for similar kind of adverse discrimination.

Deontological approaches emphasize respect of individuals’ liberty and rights or conformance with rules as the hallmark of justice. Rights-based approaches associate justice with unconstrained liberty and private property rights, which are argued to offer equality of opportunity. However, they ignore unequal capacities to exercise liberty and property rights. Moreover, they do not deal well with unequal starting points. Locke qualified private property by saying that owners should leave enough of same quality for others. This amounts to denial of scarcity and would not be applicable to situations where parties have mutually incompatible interests. Other approaches legitimate starting points by postulating a voluntary social contract among equals. Rights-based approaches emphasize that rights ought to be respected even if doing so would prevent the achievement of consequences such as utility maximization or equality of some kind. Rights-based approaches support rights to healthy environment and physical integrity. These kinds of rights have been claimed by environmental justice movements in addition to demands for equality. Environmental rights, such as rights to healthy environment have also been codified in dozens of national constitutions. In South Africa, constitutional environmental rights have been appealed to in court cases (McDonald 2002).

The above discussion has mainly focused on intragenerational distributive justice. However, it can also be applied to procedural justice which encompasses recognition, participation and the distribution of power in environmental decision-making (Fraser & Honneth 2003; Lind & Tyler 1988). For the utilitarian approach, decision-making processes ought to serve the maximization of

utility. This is why it often suggest market-like arrangements which treat people as having similar abilities to protect their interests and which distribute decision-making power according to ability and willingness to pay. Deontological approaches can justify the recognition and consultation of affected parties but it has difficulties in taking the differences in their starting points into consideration. This does not pose a difficulty for other consequentialist approaches which can justify reverse discrimination on the basis of, for example, need as discussed above (Schrader-Frechette 2002).

All approaches to environmental justice raise issues that have been considered of importance by the environmental justice movement (Schlosberg 1999; Schrader-Frechette 2002). Utilitarian and welfarist approaches justify redistributions which improve utility and welfare and remind that deviations from equality of opportunity or equality of means are not always unfair. Other consequentialist approaches are helpful because they can justify actions that seek to rectify background inequalities. Deontological approaches in turn underline the importance of recognition, participation and fair distribution of power as important aspects of procedural justice. Therefore, environmental justice is best approached from a pluralist perspective rather than from the viewpoint of any one of the above discussed approaches (Schlosberg 1999). The following section suggests that the emerging environmental justice policies are indeed informed several perspectives on environmental justice.

Environmental Justice Policies

Environmental justice policies have somewhat longer history than is usually recognized. For example, the US National Environmental Policy Act (NEPA) of 1969 introduced the requirement for the

preparation of Environmental Impact Statements (EIS) for federal proposals or actions significantly affecting the human environment. NEPA required that relevant federal, state and local government agencies should be consulted in the preparation of EIS and that the EIS should be communicated to the public in hearings organized in accordance with the Administrative Procedures Act. These provisions relate to procedural environmental justice by recognizing a right to know and by enhancing public participation in environmental decision-making. Yet NEPA was not a big step forward in environmental justice as even an adverse EIS would not necessarily prevent the adoption of a project. Gradual relaxation of rules of standing by the US Supreme Court in judicial review of administrative decision-making in the 1960s and early 1970s (Orren 1976) was probably more consequential.

NEPA provided the obvious model for environmental impact assessment legislation elsewhere as well. For example, the European Union adopted a directive on the assessment of the effects of public and private projects on the environment (85/337/EEC). The directive required the member states to enact national legislation in accordance with its provisions. Article 6 of the directive required public consultation in the environmental impact assessment process although it left the choice of consultation processes and practices to national legislators. European Union's directive (96/61/EC) on integrated pollution prevention and control tied the environmental impact assessment to awarding of permits to polluting facilities. Article 6 of this directive provided that new or modified facilities could attach their EIA documents to their permit application and Article 9 provided that the EIA should be considered in the granting of permits. Article 15 in turn provided that permit applications must be made available for the public in good time for comment and

that the decisions on permits must also be made publicly available. The European EIA legislation may thus have been more consequential for procedural environmental justice than the US one.

Obviously, NEPA was not the only piece of federal legislation concerning procedural justice in environmental matters in the United States. Administrative Procedures Act of 1946 lays down the basic rules for making rules and decisions in all federal administrative agencies and states have similar legislation. In informal rule making process, administrative agencies are required to publish proposed rules in *Federal Register* and to provide an opportunity for the public to comment in writing. The final rules must also be published, with a statement explaining and justifying the rule in the light of public comments. The provisions for formal rulemaking require formal public hearings. The agency has to make public its findings of fact, conclusions of law and its decision on the rule. Elements of both procedures are combined in hybrid rule making. APA also lays down the basic rules for adjudicating – for example for granting permits – and provides for judicial review of rules and adjudications. Interpretation of standing governs who can challenge administrative rules and adjudications in judicial review. Amendments of APA such as the Freedom of Information Act of 1966 and Government in Sunshine Act of 1976 have provided for better access to information. Many federal environmental policies such as the Clean Air Act (1971) and Clean Water Act (1972) also contain provisions for citizen suits that addressed concerns for both distributive and procedural environmental justice.

Procedural environmental justice has also received more attention elsewhere. The Aarhus Convention on Access to Information, Public Participation in Decision-Making and Access to Justice in Environmental Matters

was agreed in Aarhus in Denmark in 1998 and came into force in 2001 (Rodenhoff 2002). Currently this regional convention has 29 parties, mainly European countries and countries in transition. The Aarhus Convention commits the parties to providing for access to information, participation and access to justice in environmental matters. Article 1 of the convention recognizes the right of every person of present and future generations to living in an environment adequate for his or her health and wellbeing. The article also commits the parties to guarantee the rights to access to information, public participation and access to justice and the other articles detail the commitments in these key areas.

The European Union has responded to the Aarhus Convention by adopting the directive (2003/35/EC) on public participation and access to justice in environmental matters. Article 2 of this directive extends public consultation and participation from project planning and the granting of permits to preparation of environmental plans and programs. The article also sets requirements for administrative procedures and provides specific guidance and requirements for the dissemination and disclosure of information. Articles 3 and 4 amend the other key directives (85/337/EEC) and (96/61/EC) to include more detailed information disclosure and dissemination requirements. These two articles also amend the EIA and integrated pollution prevention and control directives to provide access to parties having a sufficient interest to a judicial or comparable review of decisions, acts or omissions to establish their substantive or procedural legality with regard to public participation provisions. All in all, the new EU provisions are similar in spirit to the more general provisions of the US Administrative Procedures Act.

Distributive environmental justice has received little attention in Europe apart from

the Aarhus Convention's endorsement of a right to an environment adequate for health and wellbeing. However, the United States has taken some initial steps in the area of distributive environmental justice. President Clinton issued the Executive Order 12898 on federal actions to address environmental justice in minority populations and low income populations in 1994. The order established the responsibility of federal agencies to identify and address the consequences of their programs, policies and actions to minority and low-income populations, to conduct research on environmental justice, and to develop agency strategies on environmental justice. The executive order also created an interagency working group on environmental justice. The Executive Order of 1994 does not go very far in addressing issues of distributive environmental justice but it has stimulated information gathering and research on incidence of environmental burdens and hazards. A dozen states have also legislated on environmental justice. They have often created advisory boards on environmental justice. Some states have also provided for research and information gathering on environmental justice issues or have mandated compliance with federal law or the consideration of environmental justice issues in brownfields analysis. Sparse case law on environmental justice is based on litigation under NEPA or state law on environmental impact assessment, or under federal civil rights legislation.

Conclusions

Environmental justice movement and empirical research informed by its concerns have drawn attention to unequal incidence of environmental burdens and hazards both in the United States and elsewhere. Philosophers, economists and other social scientist are also increasingly recognizing the

concerns for distributive and procedural justice in environmental matters. Therefore, it is likely that public policy will be increasingly sensitive to environmental justice concerns in the future. The first steps have already been taken both in the United States and in Europe to incorporate environmental justice concerns into environmental governance. Most efforts have related to improving access to information, participation in environmental decisions, and access to justice which are all important concerns for procedural environmental justice. Distributive environmental justice has been recognized in the United States as a policy issue while in Europe it has not yet resulted in policy responses. However, even in the United States existing measures merely provide for information gathering. Yet it is foreseeable that increasing emphasis will be based on the assessment of justice implications of environmental plans and decisions, on the tailoring of governance solutions so as to achieve racially, spatially, and socio-economically just incidence of environmental burdens, and on the fair involvement and influence of affected parties in environmental planning, decision-making and management.

Selected References

- Agyeman, J.; R.D. Bullard and B. Evans. (2003) *Just Sustainabilities: Development in an Unequal World*. London: Earthscan Publishers.
- Adger, W.N. Huq, S. Mace, M.J. Paavola, J. eds (2005) *Fairness in Adapting to Climate Change*. Cambridge, MA: MIT Press.
- Anand, R. (2004) *International Environmental Justice: A North-South Dimension*. Aldershot: Ashgate Publishing.
- Attfield, R. (1999) *The Ethics of the Global Environment*. Edinburgh: Edinburgh University Press.

- Barry, B. (1999) "Sustainability and Intergenerational Justice", in Dobson, A. (ed.), *Fairness and Futurity: Essays in Environmental Sustainability and Social Justice*. Oxford: Oxford University Press, 93-117.
- Brainard, J.S.; A.P. Jones; I.J. Bateman; A.A. Lowett and P.J. Fallon. (2002) "Modelling Environmental Equity: Access to Air Quality in Birmingham, England", *Environment and Planning A*, 34, 695-716.
- Brockington, D. (2002) *Fortress Conservation: The Preservation of the Mkomazi Game Reserve, Tanzania*. Bloomington: Indiana University Press.
- Bryant, B. and P. Mohai. (1992), *Race and the Incidence of Environmental Hazards*. Boulder, CO: Westview Press.
- Bullard, R. (1990) *Dumping in Dixie: Race, Class and Environmental Quality*. Boulder, CO: Westview Press.
- Bullard, R.D. (1999) "Dismantling Environmental Racism in the USA", *Local Environment*, 4, 5-19.
- Ciriacy-Wantrup, S.V. (1968) *Resource Conservation: Economics and Policies*. Third Edition. Berkeley: University of California, Division of Agricultural Science.
- Denoon, D. (2000) *Getting under the Skin: The Bougainville Copper Agreement and the Creation of the Panguna Mine*. Melbourne: Melbourne University Press.
- Dobson, A. (1998) *Justice and the Environment: Conceptions of Environmental Sustainability and Dimensions of Social Justice*. Oxford: Oxford University Press.
- Dobson, A. (1999) (Editor) *Fairness and Futurity: Essays in Environmental Sustainability and Social Justice*. Oxford: Oxford University Press.
- Ekins, P.; S. Simon; L. Deutch; C. Folke and R. De Groot. (2003) "A Framework for the Practical Application of the Concepts of Critical Natural Capital and Strong Sustainability", *Ecological Economics*, 44, 165-85.
- Farmer, M.C. and A. Randall. (1998) "The Rationality of Safe Minimum Standard", *Land Economics*, 74, 287-302.
- Fraser, N. and A. Honneth. (2003) (Editors) *Redistribution or Recognition? A Political-Philosophical Exchange*. London: Verso Books.
- Gleeson, B. and N. Low. (2001) (Editors) *Governing for the Environment: Global Problems, Ethics and Democracy*. Basingstoke: Palgrave.
- Institute of Medicine (1999) *Toward Environmental Justice: Research, Education and Health Policy Needs*. Washington DC: National Academies Press.
- Klinenberg, E. (2002) *Heat Wave: A Social Autopsy of Disaster in Chicago*. Chicago: Chicago University Press.
- Kolm, S.-C. (1996) *Modern Theories of Justice*. Cambridge, MA: MIT University Press.
- Larson, B.A.; S. Avaliani; A. Golub; S. Rosen; D. Shaposhnikov; E. Strukova; J.R. Vincent and S.K. Wolff. (1999) "The Economics of Air Pollution Health Risks in Russia: A Case Study of Volgograd", *World Development*, 27, 1803-19.
- Lind, E.A. and T.R. Tyler. (1988) *The Social Psychology of Procedural Justice*. New York & London: Plenum Press.
- Little, I.M.D. (2002) *Ethics, Economics and Politics: Principles of Public Policy*. Oxford: Oxford University Press.
- Low, N. and B. Gleeson. (1998) (Editors) *Justice, Society and Nature: Exploration of Political Ecology*. Routledge: London.
- McDonald, D.A. (2002) (Editor) *Environmental Justice in South Africa*. Athens, OH: Ohio University Press.
- Meeker, E. (1974) "The Social Rate of Return on Investment in Public Health, 1880-

- 1910", *Journal of Economic History*, 34, 392-419.
- Miller, D. (1999) "Social Justice and Environmental Goods", in A. Dobson (Editor), *Fairness and Futurity: Essays in Environmental Sustainability and Social Justice*. Oxford: Oxford University Press, 151-72.
- Mitchell, G. Dorling, D. (2003) "An Environmental Justice Analysis of British Air Quality", *Environment and Planning A*, 35, 909-29.
- Morello-Frosch, R. Pastor Jr, M. Porras, C. Sadd, J. (2002) "Environmental Justice and Regional Inequality in Southern California: Implications for Future Research", *Environmental Health Perspectives Supplement*, 110, 149-54.
- Mpanya, M. (1992) "The Dumping of Toxic Waste in African Countries: A Case of Poverty and Racism", in Bryant, B. Mohai, P. eds., *Race and the Incidence of Environmental Hazards*. Boulder, CO: Westview Press, 204-14.
- Neuzil, M. Kovarik, W. (1996) *Mass Media and Environmental Conflict: America's Green Crusades*. Thousand Oaks, CA: Sage Publications.
- Neumann, R.D. (1998) *Imposing Wilderness: Struggles over Livelihood and Nature Conservation in Africa*. Berkeley: University of California Press.
- Norton, B. (2002) "The Ignorance Argument: What Must We Know to Be Fair to the Future?", in Bromley, D.W. Paavola, J. eds. *Economics, Ethics and Environmental Policy: Contested Choices*. Malden, MA: Blackwell, 35-52.
- Orren, K. (1976) "Standing to Sue: Interest Group Conflict in the Federal Courts." *American Political Science Review*, 70, 723-41.
- Paavola, J. (2004) "Law on Water Pollution and Air Pollution", in S. Krech III, J.R. McNeill and C. Merchant (Editors), *The Encyclopedia of World Environmental History*. London & New York: Routledge.
- Paavola, J. and I. Lowe. (2005) *Environmental Values in the Globalizing World: Nature, Justice and Governance*. London & New York: Routledge.
- Pastor Jr, M. Sadd, J.L. Morello-Frosch, R. (2004) "Reading, Writing, and Toxics: Children's Health, Academic Performance, and Environmental Justice in Los Angeles", *Environment and Planning C*, 22, 271-90.
- Perfecto, I. (1992) "Pesticide Exposure of Farm Workers and the International Connection", in Bryant, B. Mohai, P. eds. *Race and the Incidence of Environmental Hazards*. Boulder, CO: Westview Press, 177-213.
- Perlin, S.A.; R.W. Setzer; J. Creason and K. Sexton. (1995) "Distribution of Industrial Air Emissions by Income and Race in the United States: An Approach Using the Toxic Release Inventory", *Environmental Science and Technology*, 29, 69-80.
- Rawls, J. (1971) *A Theory of Justice*. Cambridge, MA: Harvard University Press.
- Sagoff, M. (1988) *The Economy of the Earth: Philosophy, Law and the Environment*. Cambridge: Cambridge University Press.
- Schlosberg, D. (1999) *Environmental Justice and the New Pluralism: The Challenge of Difference for Environmentalism*. Oxford: Oxford University Press.
- Shiva, V. (1999) "Ecological balance in an era of globalization", in N. Low (Editor), *Global Ethics and Environment*. London & New York: Routledge, 47-69.
- Shrader-Frechette, K. (2002) *Environmental Justice: Creating Equality, Reclaiming Democracy*. Oxford: Oxford University Press.
- Steinberg, T. (1991) *Nature Incorporated: Industrialization and the Waters of New England*. Cambridge: Cambridge University Press.

- Stephens, C. Bullock, S. Scott, A. (2001) *Environmental Justice: Rights and Means to A Healthy Environment for All*. Special Briefing Paper 7, WSRC Global Environmental Change Program.
- Stradling, D. (1999) *Smokestacks and Progressives: Environmentalists, Engineers, and Air Quality in America, 1881-1951*. Baltimore & London: Johns Hopkins Press.
- Szreter, S. Mooney, G. (1998) "Urbanization, Mortality, and the Standard of Living Debate: New Estimates of the Expectation of Life at Birth in Nineteenth-Century British Cities", *Journal of Economic History*, 51, 84-112.
- United Church of Christ Commission on Racial Justice (1987) *Toxic Wastes and Race in the United States*. New York: United Church of Christ.
- US Environmental Protection Agency (EPA) (1992) *Environmental Equity: Reducing Risk for All Communities: I & II*. Washington DC: Environmental Protection Agency.
- US General Accounting Office (GAO) (1983) *Siting of Hazardous Waste Landfills and Their Correlation with Racial and Economic Status of Surrounding Communities*. Washington DC: USGAO. GAO/RCED-83-168.
- US President (1992) *Executive Order 12898: Federal Actions to Address Environmental Justice in Minority Populations and Low Income Populations*. Federal Register, 59.32, February 16.
- Walzer, M. (1983) *Spheres of Justice: A Defence of Pluralism and Equality*. Oxford: Blackwell.

Jouni Paavola
 School of Earth and Environment
 University of Leeds, Leeds, UK
 j.paavola@leeds.ac.uk

Family Law and Family Court

Aspasia Tsaoussis

Introduction

Family courts handle cases where parties seek court intervention to solve their family issues. Judges in family courts typically hear and decide cases involving divorce, child custody, support and visitation, access to children (or, to use the new term, “contact concerning children”), occupancy of the family residence, paternity, child abuse and neglect, juvenile delinquency and persons in need of supervision, foster care approval and review, termination of parental rights, adoption and guardianship. The subject matter of the family court is very broad, encompassing many areas of social life. However, the prominence of family-related social problems has rendered the task of the family court judges more arduous and complex. Alarming trends like the rising divorce rate, the high incidence of domestic violence and child abuse, together with the feminization and infantilization of poverty have incited strong scholarly and empirical interest in the family, producing a now voluminous body of knowledge. At the same time, new developments spurred on by advances in medical technology like in vitro fertilization and other methods of assisted reproduction are feeding the family courts with a new set of problems (e.g. relating to surrogacy).

Typically, the family court also provides an array of other “in-house” services: domestic abuse service centers, divorce education and counselling centers, guardian ad litem programs, and conciliation centers, to name only a few. Thus, each family court has its own “support system”, made up of a wider network of professionals who have been trained in other disciplines and contribute their expertise in view of facilitating the court process, e.g.

psychologists, public servants, social workers, conciliators, mediators, and social services employees. Although the variety of services differs from each country to the next, the basic administrative structure is the one previously described.

Comparative Guide to Family Court System

Normal courtroom checks and balances do not exist in family court. Unlike criminal and civil courts, family courts have no juries. They also do not mandate legal representation. For example, in Canada, lay persons can easily complete applications in handwriting. Therefore, the only litigants with attorneys are those who can afford them. In most legal orders, family court judges have broad discretionary powers. It is not uncommon, for example, for judges to hold hearings in which important rulings are made with only one party present (*ex parte* hearings). Such hearings can violate basic constitutional rights of procedural due process. Hearings must be serious and lengthy enough to allow a genuine review of the issues and actions before the court. There are obviously no hard and fast rules, but for example, in the American family court system, the National Council of Juvenile and Family Court Judges has produced a detailed analysis of what issues should be covered in the most common hearings: preliminary protective, adjudicatory disposition, review, permanency planning, termination of parental rights, and adoption (National Council of Juvenile and Family Court Judges 1995:29-105).

There are some variations in the structure and the workings of the family court, even among countries with advanced family court systems such as the U.S., Australia, Japan, Canada, and several countries in Western Europe. For example, in the US, the jurisdiction of family courts differs somewhat

from state to state. From November 1996 through June 1999, the American Bar Association (ABA) developed six Unified Family Court (UFC) systems in three U.S. states and one territory and created a network of national groups to help educate the public about Unified Family Courts (see esp. Babb 1998). In England, the High Court has a Family Division with jurisdiction over family matters. Australia has a national specialist court, the Family Court of Australia, which was established to implement the gender-neutral philosophy of the Family Law Act 1975 (Commonwealth).

Continental legal systems have been strongly influenced by the German-Roman tradition. Therefore, their family court system is characterized by a heavier dependence on statutory law. Some European countries (e.g. Russia, Greece and Malta) have no separate family courts, but divisions or sections of civil courts charged solely with the adjudication of family disputes. The judges who sit in “family court” are appointed on a rotating basis and their term typically lasts three years. Finally, there are countries in Eastern Europe (like Hungary) that are currently in the process of creating so-called “family law groups” within their courts.

The Effectiveness of Family Courts

General dissatisfaction has been expressed by theorists, practitioners, and family litigants about the fact that the court system is costly, complex, and sometimes insensitive to the litigants’ needs. There is also criticism that the sheer volume of the family court workload results in less than perfect outcomes for many. In most countries, family courts are overburdened with cases and do not fare well in handling the coordination problems that the lengthy trials of certain cases demand (e.g. child-abuse cases that entail multiple appearances before the court and repeated interviews with the children). The

considerable delays in the processes of the family court have also come under censure (see e.g. Star 1996) and are often discussed under the prism of the courts’ severe budgetary constraints. As the U.S. National Council of Juvenile and Family Court Judges (1995:10) explains:

“Unfortunately, many courts have neither the ability nor the resources to meet these new demands. Judicial caseloads have actually risen at the same time that the number of issues, hearings, and parties has increased. As a result, in many jurisdictions, the quality of the court process has gravely suffered. Hearings are often rushed in child abuse and neglect cases. There are also frequent and unfortunate delays in the timing of hearings and decisions, causing children to grow up without permanent homes. Many courts know little about relevant agency operations or services. All too often, child welfare agency employees spend unnecessary hours waiting for court hearings while they could be “out working in the field.”

Family courts adjudicate a wide range of cases. Concomitantly, the related policy issues that arise are complex and multifaceted. Reviewing the literature, one can identify a great divide between countries that have a long-standing tradition in family court practice and those that do not. It is no coincidence that the former also have more systematized methods of collecting and processing valuable information (e.g. from the annual reports issued by family courts that allow policymakers to track changes over time and use them as indicators of effectiveness). However, it is possible to pinpoint certain areas of practice that have attracted the most attention from theorists and practitioners across the globe. These areas will be discussed in greater detail, by

reference to the empirical evidence that is currently available.

Safeguarding the Interests of Children

The “best interests of the child standard” carries particular legal, moral and political weight in divorce proceedings. An extensive body of empirical literature now exists relating to the implementation of this standard and more generally to child welfare law after the United Nations Convention on the Rights of the Child, which affirmed that the standard will be the basic concern of parents and guardians in rearing children [see article 18(1) and also article 9(3)]. Because the standard “is more a vague platitude than a legal or scientific standard, it is subject to abuse both by judges who administer it and parents who use it to further their own interests” (Charlow 1994:3). Since it is not defined anywhere in statute or case law, its particular content is determined by the court in each different case, with the assistance of a child advocate or guardian ad litem, who is the court-appointed protector of a child’s best interests.

In the United States, a *guardian ad litem* may be appointed by the court to submit to the court an objective statement of fact after having undertaken an investigation into the case so that the court can better ascertain what may be in the best interests of the child. The duty of the guardian ad litem is to interview the minor and all other individuals relevant to the case, to review pertinent medical and school records, and to report back to the court the results of this investigation (see ABA Concept Paper on Family Law 1999). Jurisdictions are split on whether guardians ad litem are charged with offering their own opinion of what course of conduct fits the best interests of the child in addition to this fact report or whether they are to report only the results of interviews and examinations. Each court is left to its own

discretion in utilizing the guardian ad litem for the protection of the children under its jurisdiction. In Sweden, legislation was introduced in 1999 whereby the court may, at the request of a prosecutor, appoint a “special representative” for a minor who is under eighteen, if the custodian is suspected of an offence for which imprisonment may be imposed or if it is feared that the custodian will not protect the best interests of the child. This court-assigned representative has broad powers in the legal proceedings, including the right to pursue an action for damages on behalf of the child.

The independence and objectivity of these child advocates or guardians are the most important determinants of how successfully they perform their role. A conscientious advocate can be of enormous value to the court, particularly when judicial resources are limited and the primary parties are unwilling, unable, or unlikely to adequately account for the child’s interests. A skillful guardian who is given clear guidelines as to his or her role can contribute greatly to the sense of fairness by participants, which is a critical component to respect for the rule of law.

On some occasions, family courts have a tough time adjusting to fast-paced reform in child welfare law. For example, there has been a recent shift in U.S. family policy toward promoting adoption (as opposed to state-managed foster care) as a solution for children in “dysfunctional” families. This policy shift was initiated by the Adoption and Safe Families Act of 1997. When the Act took effect in 1998, it was followed by sweeping amendments of child welfare legislation at the State level.

In Australia, the Family Law Act 1975 provides that, subject to any contrary orders, both parents are guardians of any child of the marriage and have joint custody. However, the Family Court has not interpreted this as a presumption in favor of joint custody where

disputed cases of custody, guardianship, or access are brought before it; instead, it is given broad powers to make orders, all of which are subject to the statutory requirement that “the court shall regard the welfare of the child as the paramount consideration”.

One proposal in the direction of best serving children’s interests is to take some authority away from the family court, like the United Kingdom did with its Child Support Act of 1991. This Act transferred the functions of assessing child support requirements in non-matrimonial settings from the judiciary to a government agency: it is this Child Support Agency that has the powers to assess the financial needs of the child, trace absent parents and recover child support money for the use of the parent or any other caretaker of the child. The Agency can enforce its orders by attaching the income of the liable parent without recourse to any judicial process. The English model is effective for a number of reasons: administrative employees are better equipped to recover child support payments and more successful at monitoring each case in all its particulars than are judges.

Finally, a greater emphasis on human rights jurisprudence is needed in order for family courts to fully recognize children’s rights under the European Convention on Human Rights. In her analysis of the relevant English case law, Fortin (2006) has argued that a reinterpretation of the “paramountcy principle” in the Children Act 1989 should be accompanied by a radically different judicial approach to evidence relating to children’s best interests.

Child Custody: New Policy Considerations

A number of policies have been proposed to ease the transition of families from marriage to divorce. Joint physical custody (or shared parenting) is one of these proposals. In many countries, there has been heated debate

concerning the establishment of a “joint custody” norm to be applied by the family courts. It is characteristic that in the U.S., over thirty-five states have passed joint custody legislation.

The merits of joint custody have been amply documented in the literature. Direct feedback from the courts remains problematic, since few family courts keep statistics regarding the custody of children. Policymakers thus rely primarily on social science research findings. Some of these findings suggest that joint custody also has its drawbacks. Social scientists who have conducted longitudinal studies on the effects of divorce on children report that joint residential custody is an arrangement that may confuse very young children (Wallerstein et al. 2000:198). Some children were found to internalize the constant back-and-forth between homes into their personalities and thus had a difficult time dealing with a stable environment.

According to Wallerstein and Blakeslee (2003), joint custody has little benefit for children because the fundamental problem is rooted in the fact that their relationship with “mom” and “dad” has forever changed. If the postdivorce relationship of the parents is harmonious, then the legal form of custody matters little. However, joint custody arrangements may be particularly harmful to children in high-conflict families. As this thirty-year study on the impact of divorce on children concludes, it is very important that parents give priority to the child’s changing capacity and need for uniform routines (Wallerstein et al. 2000, pp. 215-216). Another empirical study supporting these conclusions finds that it is not the amount of time that non-resident fathers spend with their children, but how they interact with their children that is important (Marsiglio et al. 2000). Researchers should continue to study fathers’ influence on their children in stable

family units and data from national surveys will remain a critical source of information about how differing levels and types of care affect children's well-being.

Another strand of the literature focuses on the parent's sexual orientation as a criterion of parental fitness in custody and visitation after divorce, as well as in adoption. Some countries have been more progressive than others in the direction of recognizing extended rights in marriage and adoption for same-sex couples. Since the late 1990s, Denmark, the Netherlands, Sweden, Iceland, and Norway have enacted "registered partnership" bills—followed by Germany, whose new Registered Partnership Act entered into force in August 2001, following heated discussions in the German Parliament. In April 2001, the first Act making the adoption of Dutch children available to same-sex couples entered into force in the Netherlands. In American family law, challenges to marriage statutes on behalf of same-sex couples, the liberalization of access to adoption records, and permitting adoption by gay and lesbian parents as joint parents are examples of recent changes at the State level.

Case law has developed in the same direction: the Supreme Court of Vermont held in 1999 that the state must extend to same-sex couples the same benefits that married couples receive; In November 2003, the Massachusetts Supreme Judicial Court held that excluding same-sex couples from the benefits of civil marriage violated the state constitution. In May of 2004 Massachusetts became the first state to legalize same-sex marriage, followed in April of 2005 by Connecticut that approved civil unions for same-sex couples. These developments suggest that future reform will establish a broader framework for the rights of homosexuals in all areas of family law and "fuel a remarkable expansion of legal shelters for lesbian and gay parent-child

relationships" (Logue 2002:129). Pressure from the gay community is also likely to rise, since same-sex couples "are in the thick of a vigorous profamily movement of their own" (Stacey 1998:177).

The legal regulation of non-marital relationships is moving towards a similar direction. It is characteristic that the Canadian courts have since the late 1990s rendered a number of important decisions that have used the Canadian Charter of Rights and Freedoms to grant greater legal recognition to same-sex partners and unmarried heterosexual cohabitants (see esp. Bala 2001). Higher courts in several other countries (e.g., in South Africa: *Fourie v. Minister of Home Affairs*, Case No. 232/2003, Supreme Court of Appeal of South Africa, November 30, 2004) are moving towards the recognition of the "family" status of same-sex unions. However, equating the rights of same-sex partners with spousal rights remains a highly controversial issue in most legal orders. Furthermore, the labeling of families as "non-traditional", or the use of other modifiers, continues to reinforce the implicit norm and definition of a "real" family (Dowd 2002:439).

Gender Stereotypes in the Family Court System

Over the past twenty years, strong criticism has been levelled against the family court system in many countries for perpetuating gender bias. According to several theorists, the court frequently not only denies the capability and desire of many men to participate actively in the upbringing of their children, but it also perpetuates the subjugation of women as mothers by deeming them weak and incapable of survival without the support of a man (see McNeely 1998).

At present, the most widely applied principle in custody disputes shows preference for the primary care-taker.

“Essentially this means confirming the *status quo* – making the parent who provided the bulk of the care-taking within marriage, the residential parent after divorce” (Richards 1994:317). In practice, this is a return to the maternal preference doctrine, which could be a reason why several feminists have strongly advocated the primary care-taker principle (see e.g. Smart and Sevenhuijsen 1989).

The fathers’ rights movement has grown as a response to concerns about discrimination against men in child custody debates, in pensions, etc. The movement has gained momentum in the U.S., in Canada (with the activism of fathers’ rights groups like the National Fatherhood Initiative), in Australia (“Men’s Confraternity”) and in Western European countries such as Germany. These groups call attention to cases where fathers have been unreasonably denied access to their children and support their arguments by reference to research demonstrating the harmful effects of a father’s absence on young children. There is also empirical evidence to attest the importance of fatherhood for men: for example, Jordan (1996) showed that separated men, particularly those whose relationships have ended against their wishes, continue to experience poor health and wellbeing outcomes up to ten years after their divorce. Fathers’ rights advocates aim to raise public awareness by emphasizing cases where the family court violated fathers’ constitutional rights. According to data collected by the American Coalition for Fathers and Children, American fathers have been increasingly filing lawsuits under the Fourteenth Amendment’s Equal Protection Clause.

In the United States, New Jersey was the first state to establish a gender bias task force with the goal of collecting valuable data on the experiences of both sexes in family courts. Since then, forty-five states and a

number of federal circuit courts have established gender bias task forces. Forty-four states have already published their final reports, showing that there is still a lot to be learned about the dynamics of judicial decision-making. In some states, like Virginia, Court Environment Committees were set up to test whether gender affected the decision-making process by having some impact on the behavior and interpersonal relationships of the players - lawyers, litigants, judges, clerks, magistrates, witnesses, jurors, and others (see the *Washington and Lee Law Review* special issue of the Summer of 2001). Resnik (1996) discusses the positive impact of task forces and their contribution in dispelling gender based myths. Task forces and entities like the National Judicial Education Program to Promote Equality of Women and Men in the Courts share the same overriding goal: educating judges.

A study commissioned in 1989 by the Supreme Judicial Court of the state of Massachusetts showed that mothers engaged in custody disputes with their ex-husbands or boyfriends can fall victim to gender bias (Abrams and Greaney 1989). Family courts held mothers to higher standards than fathers. Judges and other courtroom personnel scrutinized mothers’ habits, work schedules, and relationships. By contrast, fathers who simply sought custody were viewed as undertaking what the study termed “an extraordinary act of commitment” to their children. The study confirmed that mothers are almost always awarded full or joint custody of their children in divorce cases where custody is not disputed. However, the study found that when there was a fight over the children, fathers won primary or joint custody more than 70 percent of the time—whether or not there was a history of spousal or child abuse.

Chesler (1986) interviewed 60 mothers involved in a custody dispute and found that fathers who contest custody are more likely than their wives to win (p. 65). In 82% of the disputed custody cases fathers achieved sole custody despite the fact that only 13% had been involved in child care activities prior to divorce (*id.*, p. 79). Moreover, 59% of fathers who won custody litigation had abused their wives, and 50% of fathers who obtained custody through private negotiations had abused their wives.

Several studies in the U.S. have linked gender bias to outcomes of custody disputes involving child-abuse claims (for an overview, see the data collected by the Leadership Council on Child Abuse and Interpersonal Violence). These studies indicate that family courts often do not consider the history of violence between the parents in making custody and visitation decisions. According to the Final Report of the Judicial Council of California's Advisory Committee on Gender Bias in the courts of California, gender bias problems are particularly acute in family courts, and most problematic when sexual abuse of children is alleged in custody or visitation proceedings (Danforth and Welling 1996). The survey showed that negative stereotypes about women encourage judges to disbelieve women's allegations of child sexual abuse. A similar report in the state of Florida concluded that many men file proceedings to contest custody as a way of forcing an advantageous property settlement (Report of the Florida Supreme Court Gender Bias Study Commission Executive Summary 1990, p. 7).

In the same vein, a Report of the American Psychological Association Presidential Task Force on Violence and the Family noted that custody litigation frequently becomes a vehicle whereby batterers attempt to extend or maintain their control and authority over the abused children

after separation (American Psychological Association 1996). The nonviolent parent may be placed at a disadvantage in the family court, and behavior that would seem reasonable as a protection from abuse may be misinterpreted as a sign of instability. Psychological evaluators not trained in domestic violence may contribute to this process by ignoring or minimizing the violence and by giving inappropriate pathological labels to women's responses to chronic victimization. The problem has escalated to such dimensions that since the late 1990s, a series of handbooks have been published (e.g. by the National Center for State Courts 1997) to provide judges and court managers guideposts for determining when domestic violence is occurring between the parties to a dispute over child custody or visitation.

Australia is a country that can boast of a rich tradition in family court practice. Much scholarship has been devoted to elucidating established or emerging principles and trends in legislative interpretation. Dickey (1997: 391-405) for example, has identified six considerations which frequently influence proceedings in parenting disputes in the Family Court of Australia. These are: (1) A preference for maintaining the status quo; (2) A propensity to not separate siblings, (3) A privileging of the mother-child relationship; (4) A privileging of natural parent relationships; (5) The wishes of the child; (6) Parental conduct.

In 1983 the Australian Institute of Family Studies (AIFS) conducted a major empirical study on the economic consequences of marriage breakdown, including custodial outcomes. Although the Institute reported a relatively high percentage of younger men who had care and control of their children, it was still overwhelmingly women who looked after their children on a full-time basis. The study concluded that "society at large still

sees the nurturing of children as being the primary responsibility of women” (Australian Institute of Family Studies 1986:268).

Another early Australian study on the outcome of custody cases (Horwill and Bordow 1983) found that fathers had a greater likelihood of being awarded custody in contested than in uncontested cases, which accounted for only 10 per cent of all orders made in the Melbourne registry of the Family Court. Although this finding suggests that the maternal preference rule was still widely applicable, it does not indicate how many fathers would personally undertake the caring responsibility and how many would delegate that responsibility to relatives, *de factos*, new spouses or hired help (Nygh 1985:67-68). A more recent analysis of a random sample of parenting judgments in the Family Court of Australia (Moloney 2000) found that traditional constructions of motherhood persist within the Family Court of Australia and that as a corollary, the nurturing role of fathers continues to be viewed with skepticism. The evidence suggests that when fathers are successful, these successes may occur largely by default.

Finally, New Zealand has a family court system that gives strong emphasis to the needs of the children and establishes a framework for decision making before the engagement of the legal system. At the center of the system there is a “family care conference” involving the parents, the children, members of the extended family, an advocate for the children, and a mediator—but no lawyers. It also lays down a list of standards that can be used to define the primary care-taker (Hassall and Maxwell 1992).

Over the last ten years, policy makers have intensified their efforts to deal with the protection of children’s rights in public and private law proceedings where they are placed at risk. Recent initiatives in the United

Kingdom best illustrate this point. In April 2001, the “Children and Family Court Advisory and Support Service” (CAFCASS) was established by the Criminal Justice and Court Services Act 2000 as a new national Non-Departmental Public Body for England and Wales. CAFCASS brought together the functions of three respected services that support vulnerable children in family proceedings: the Family Court Welfare Service; the work of the guardians ad litem and Reporting Officers; and the Children’s Divisions of the Official Solicitor. CAFCASS’ primary duties are to safeguard and promote the welfare of the child and to give advice to any court about any application made to it in such proceedings. In March 2002, a Lord Chancellor’s Department Consultation Paper recommended an overarching Family Justice Council as an effective mechanism for coordinating inter-agency working within the family justice system. In 2004, the Family Justice Council for England and Wales was established. Its Chair is the President of the Family Division of the High Court.

Self-Representation:

A New Direction for the Family Court?

Self-representation is a recently burgeoning trend in family court practice, especially in Australia (see characteristically the third report of the Senate’s Legal and Constitutional References Committee 1998) and the United States—to make legal aid available to a larger number of litigants (). Policy makers are increasingly interested in the impact that this development is having on legal service providers and the administration of justice generally.

In the United States, the pro se litigation movement sprung forth in the early 1990s, following the publication of statistics from a national state courts study indicating that the number of self-represented litigants varies

from 30–80 percent depending on the state, while Australian studies show that the rate in this country is approximately 37% (Cox 2002). Three researchers with funding from the American Bar Association released the results of the first empirical study of self-representation in a civil-court setting after tracking a year's worth of domestic relations cases in Phoenix. The survey unearthed a startling statistic: in 88 percent of divorce cases in Maricopa County Superior Court, at least one litigant was self-represented (see Sales et al. 1993).

In recent years, South Africa (a country that does not have a distinct family court) has taken steps to empower family law litigants in person, with the initiatives of non-governmental organizations working closely with the Department of Justice and Constitutional Development. The model of self-representation developed by a particular not-for-profit organization that provides family law services is quite interesting (see Baartman 2002), because it took into account the harsh realities of this country: over 85% of all people before the court are unrepresented and a high percentage of the general population are functionally illiterate and innumerate.

Despite its merits, self-representation raises serious concerns of accessibility for the average litigant: if the family court represents a system of justice administering a body of law whose core essence is a bundle of fundamental human rights, how can citizens of less privileged educational and/or economic backgrounds find justice? If they lack the knowledge to identify where their best interests lie, they cannot be adequately self-represented. Given the sometimes prohibitive costs of legal representation, it seems that self-representation could add yet another obstacle – or even worse, a barrier – to many citizens' access to the family court. Finally, geographical barriers also divide

litigants along lines of accessibility to the courts, so much so that a system of “traveling justice” (Rasnow 2002) is needed to ensure that limited access communities are provided pro se assistance. On the other hand, being self-represented is arguably a better alternative to being unrepresented.

Mediation for Greater Efficiency

Family litigation tends to be very contentious and incredibly acrimonious. This can lead to very poor outcomes for those litigants the family court most intends to protect. For example, litigated custody disputes rarely attain the primary objective of producing results in the best interests of the children. Many countries are striving to address this problem by steering away from the adversarial court process and by devoting more resources, such as trained family law counselors, to the mediation of family disputes. A turn to mediation is an important structural change presupposing that judges, lawyers and litigants adopt a mindset that contrasts sharply with the one that the traditional adversarial paradigm rests upon.

The consensual resolution of family disputes is swifter, less expensive (see esp. Kelly 1990) and produces more durable agreements compared to litigation. Empirical studies show that agreements reached by the parties after mediation are honored far more effectively than any decisions imposed by the court, making mediation the most favored alternative in family law practice (see Mosten 1997 and Boulle et al. 1998). The process of mediation facilitates communication between the parties and permits the voluntary exchange of information that constitutes the basis for any mutually beneficial agreement. More importantly, when children are involved, non-adversarial proceedings reduce the stress and trauma that children customarily experience in family court hearings. Finally, mediation can protect the

parties from the uncertainty of judicial outcomes that are not predictable or stable (Moloney 2000).

In what concerns disputes arising from divorce, mediation has become a popular and frequently used alternative to adversarial proceedings across the United States over the last 20 years. While the promise of mediation has not always been fully realized, the benefits have been substantial enough to have led 38 states and the District of Columbia to develop programs within their court systems to mediate custody, visitation, and child support (Keilitz et al. 1992). In addition to these court-based or court-annexed programs, private divorce mediation is a growing alternative to adversarial divorce as evidenced by the more than tripling of membership in the past 10 years in organizations such as the Academy of Family Mediators (Center for Divorce Mediation and Alternative Dispute Resolution: Study of Divorce Outcomes 2006). In contrast to many court-based programs, private mediation addresses all divorce issues. The feedback from court-based programs suggests that mediation is perceived as “better” than the standard litigation process both in terms of the level of litigant satisfaction and compliance with agreements. Research confirms that mediated divorces take less time than adversarial divorces and are significantly less likely to result in post-judgment modification, thus sparing couples and families added emotional and financial costs (see esp. Marcus et al. 1999).

In Canada, virtually every province funds and operates some form of family mediation program. These programs are either directly connected to the courts, or are community-based with links to the courts and other social services. Mediation is not mandatory, however contentious cases proceed to trial only after mediation has been attempted, conducted by family court counselors.

Australia, a pioneer in family court policy reform, sought to improve access to Primary Dispute Resolution (PDR) services for families in 1997, only two years after the enactment of the Family Law Reform Act of 1995. The Australian government embarked on a process of examining how to effect structural changes that would offer the greatest opportunity for people to resolve their family disputes without recourse to litigation. A principle reformatory goal was that the PDR services should focus on the needs of clients and provide equity of access. One of its primary concerns was how to utilise the expertise in court processes of the current court counsellors and mediators in a new structure developed in the community sector. Australia has such a successful track record of family law reform that it could well act as a model for many countries. The Australian Government has recently made a sizeable investment (amounting to \$397 million over four years) in the family law system. This includes the establishment of 65 Family Relationship Centres and a national advice line.

In Europe, several jurisdictions have well-established systems of family mediation. Since the early 1990s there exists a European Charter for the Training of Family Mediators in Situations of Divorce and Separation. In the U.K., the National Family Mediation is a network of over 60 local not-for-profit Family Mediation Services in England and Wales offering help to couples, married or unmarried who are in the process of separation and divorce. In Germany, whose legal order had been principally geared towards litigation and arbitration, general interest in non-adversarial ADR methods was relatively modest until the late 1990s. In 2000, the Federal Government of Germany introduced legislation permitting all German states to introduce mandatory court-connected mediation with respect to certain kinds of

civil disputes. The mandatory mediation requirement does not apply to family disputes. The German system is more legalistic relative to the American or Australian one.

In France, family mediation was introduced from Quebec in the late 1980s, and seen as a panacea for dealing with the rapidly increasing divorce rate and enormous backlogs in the courts. However, French government, legal institutions, law, culture, religion, and family structures have exerted their influence on the path family mediation has taken subsequently in France to produce family mediation with its own distinctive “Latin logic” (MacFarlane 2004). In Portugal, the Family Mediation Office is responsible not only for providing guidance, conducting investigations, disseminating information and providing training, but also, through mediation, for providing parents who are in the process of separation and/or divorce with a context for negotiation, ensuring the relationship between parents and children continues, fostering co-parenting, helping agreements concerning the children to be reached, and facilitating communication between the parents.

In Japan, court-connected mediation (*“chotei”*) is based on an agreement between the parties that is facilitated by the intervention of a summary or a district court. Mediation is mandatory for all family cases. Unless otherwise provided, general jurisdiction for all non-family disputes that go to court-connected mediation is in one of the summary courts located throughout Japan. The central feature of the Japanese system is to direct the parties towards a settlement of dispute consistent with reason and befitting actual circumstances “by mutual concession” (see article 1 of the Civil Conciliation Act). In mainstream alternative dispute resolution theory, mutual concession (or compromise) is less desirable than a problem-solving solution

that satisfies the underlying interests of both parties. The main reason mutual concession is specifically provided for by law is that it reflects the cultural tradition of consensus in Japan (Funken 2002).

In India, the Family Courts Act of 1984 provided for the establishment of Family Courts “with a view to promote conciliation in, and secure speedy settlement of disputes relating to marriage and family affairs and for matters connected therewith”. In Bangladesh, a former colonial state that has developed a legal system largely based on English common law, alternative dispute resolution in its rudimentary form is applied to increase accessibility to justice. Because of the difficulty in accessing the courts and the widespread corruption (according to one independent sample survey conducted by Transparency International Bangladesh, over 60 percent of the persons involved in court cases paid bribes to court officials because litigation is time consuming), alternative dispute resolution by traditional village leaders is popular in rural communities.

In countries such as Malta which enacted family court legislation more recently, the debate centered on whether counseling services and conciliation should be mandatory prior to access to court proceedings (Farrugia 2001:286). The opinion which prevailed was to encourage spouses to seek marital therapy and to make access to such conciliation services easier. The experience from Northern Ireland indicates that the information meeting (the compulsory precursor for the initiation of divorce proceedings under the Family Law Act), is not fulfilling its objective to help savable marriages and to promote mediation as an alternative to litigation (Glennon 2001:353). In the latest amendment of its marriage law, Austria also introduced mediation in the divorce procedure. The new provisions emphasize the mediator’s

obligation of secrecy and call for criminal sanctions if this duty is violated. In Cyprus, court-annexed mediation is the object of debate, with the pending Family Disputes Mediation Law (see Cyprus Mediation Association 2006). Finally, the Baltic countries (Estonia, Latvia and Lithuania) have also developed independent conflict management centers that provide community-based mediation services to resolve a range of local neighborhood, family and business disputes. Their main goals are to pursue legislation incorporating mediation as a formal component of the justice system, and to facilitate the empowerment of under-represented and disadvantaged groups.

In the literature, one finds concerns that court-connected mediation procedures may hurt women. Empirical evidence suggests that women generally fare worse than men in mandatory mediation. In her critique of California's mandatory mediation system for disputed child custody issues, Grillo (1991) concludes that when mediation is imposed rather than voluntarily engaged in, its virtues are lost. Bryan (1992 and 1997) identifies the risks of women's participation in divorce mediation as a problem of power imbalance between men and women. Yet several empirical studies show that those in mediation had more positive perceptions of their divorce experience. Comparing attitudes of couples who had mediated all their divorce issues (and not just custody issues), Kelly found that women in mediation were equally satisfied as men with the overall process, mediator impartiality, adequacy and clarity of data, and various mediator techniques (Kelly 1989:84-86). Pearson (1991) conducted interviews with over 300 people who mediated at least some of the issues in their divorce, finding that mediation is not worse than adversarial decision-making in generating agreements that are perceived to be equitable and fair (*id.* at 192-193).

Perhaps it is Brinig (1995:34) who expressed it best: "[C]ongested courts cannot justify mandatory mediation in cases where one spouse holds a monopoly on marital power. No one should order mediation when there has been abuse within the family, substance abuse, or systematic hiding of assets". Battered women and children may be hurt or endangered in custody mediation [see Hart (1990); see also Astor (1994), discussing the Australian experience from handling violence in family mediation]. In a review of the literature and an analysis of data from four publicly-funded Canadian Mediation Programs, Goundry et al. (1998) conclude that serious caveats still exist about the inability of mediation services to screen out, with a high degree of accuracy, those women who have experienced violence or abuse and for whom mediation could provoke a dangerous situation.

Further analysis is required to dispel the concerns about gender bias in mediation and to understand the broader implications of gender differences in all forums where family disputes are resolved. However, there seems to be general consensus concerning the importance of having well-functioning mechanisms of non-judicial redress for litigants wishing to resolve their family-related disputes in an amicable manner. The courts adopting these new mechanisms should be well aware of both their potential benefits and their potential costs.

Conclusion

No unifying coherent family policy can exist without a well-functioning, viable family court system. Perhaps the greatest advantage of the family court is that it offers judges significant training in resolving family-related disputes. Indeed, judges today are more enlightened about particular social problems originating in the family (like domestic violence and child sexual abuse) than judges

were two decades ago. Many family-court judges spent legal careers practicing domestic relations and child welfare law. Undoubtedly, their knowledge of, and sensitivity to all related issues has grown. The newly emergent trend toward family court unification elevates the judge to key player status, as the central idea of the unified family court is a single and highly specialized judge who hears the family's multiple cases under a comprehensive jurisdiction. At the same time, the experience from the unified family court system demonstrates the pivotal role of the multidisciplinary team of therapeutic and dispute resolution professionals who make recommendations to the judge and provide support, and in a very real sense, "therapeutic justice" (Kuhn 1998:91) to children and families.

The educational role of the family court is increasingly important in these times of fluidity of family forms. In the past twenty years, non-traditional families have increased in number, posing challenges to legislators, policy makers and judges. Unified family courts routinely encounter what can be defined as a "postnuclear family" (Petre 1999:161), whose family members are generally identifiable only through matrilineal descent. that is, a family This is to a certain extent safeguarded by the framework laid down by international private family law, which continues to evolve in the direction of becoming more child-centered and gender-sensitive (see the two additional Protocols to the Convention on the Rights of the Child and the Optional Protocol to the UN Convention on the Elimination of All Forms of Discrimination against Women, both introduced by the UN General Assembly in 2000).

Family courts adjudicate within this global framework of human rights (see Nicholson and Harrison 2000), but it seems that they should go one step further by remaining

attuned to the special needs of socially disadvantaged groups. Women remain politically and socially disempowered in most developing nations of the world. In highly industrialized countries, women have long enjoyed the benefits of formal equality, yet asymmetrical gender roles within the household continue to burden them with child care and housework responsibilities, as attested by a plethora of empirical studies. In the Beijing Plus 5 Annual Summit for Women (2000), the UN General Assembly acknowledged that the structure of the family does not always provide adequate support for women and that this undermines efforts to achieve substantial gender equality.

Furthermore, if the family court is to embody the norms and values of democratic and open societies, it should not only acknowledge the changes in the perceptions of family life, the changing content of gender roles and the new ethical dilemmas posed by medical interventions in the family sphere. The family court should develop a case law that recasts, redefines, and restructures. Cutting through the social fabric, it has the potential to shape the ideas, norms and values that will predominate in the legal culture of the first half of the twenty-first century.

Selected References

- Abrams, Ruth I. and John M. Greaney. (1989) *Report of the Gender Bias Study of the Supreme Judicial Court, Commonwealth of Massachusetts*. Boston, MA: The Court.
- American Bar Association, Central and East European Law Initiative. (1999) *Concept Paper on Family Law*. www.abanet.org/ceeli/nosearch/archive/conceptpapers/familylaw/familylaw.html
- American Psychological Association. (1996) *Report of the American Psychological Association Presidential Task Force on Violence and the Family*. Washington, DC. Available at:

- www.apa.org/pi/pii/familyvio/issue5.html.
- Astor, Hilary. (1994) "Violence and Family Mediation Policy", *Australian Journal of Family Law*, 8, 3-21.
- Australia, Attorney General's Department. (1997) *The Delivery of Primary Dispute Resolution Services in Family Law*.
- Australian Institute of Family Studies. (AIFS). (1986) *Setting Up: Property and Income Distribution on Divorce in Australia*. Melbourne: Prentice-Hall.
- Babb, Barbara A. (1998) "Where We Stand: An Analysis of America's Family Law Adjudicatory Systems and the Mandate to Establish Unified Family Courts", *Family Law Quarterly*, 32.1, 31-65.
- Bala, Nicholas. (2001) "Court Decisions on Same-Sex and Unmarried Partners, Spousal Rights and Children", Andrew Bainham ed., *The International Survey of Family Law. (2001 Edition)*, 43-63. Bristol, U.K.: Jordan Publishing.
- Baartman, Elizabeth. (2002) "Reflections on a South African Initiative: Empowering Family Law Litigants in Person". www.pflc.org.za/press0211_speechmelbourne.htm.
- Bordow, Sophy and Gibson, J. (1994) "Evaluation of the Family Court Mediation Service", Research Report 12, Family Court of Australia Research and Evaluation Unit.
- Boulle, Laurence, Judi Jones and Virginia Goldblatt. (1998) *Mediation: Principles, Process, Practice*. Wellington, N.Z.: Butterworths.
- Brinig, Margaret F. (1995) "Does Mediation Systematically Disadvantage Women?", *William and Mary Journal of Women and the Law*, 2, 1-34.
- Bryan, Penelope E. (1992) "Killing Us Softly: Divorce Mediation and the Politics of Power", *Buffalo Law Review*, 40, 441-523.
- Bryan, Penelope E. (1997) "The Coercion of Women in Divorce Settlement Negotiations", *Denver University Law Review*, 74, 931-940.
- Center for Divorce Mediation and Alternative Dispute Resolution: *Study of Divorce Outcomes*. Norwalk, CT: CDMADR. www.divorcemediation.norwalk.ct.us/study_of_divorce_outcomes.htm
- Charlow, Andrea. (1994) "Awarding Custody: The Best Interests of the Child and Other Fictions", in S. Randall Humm, Beate Anna Ort, Martin Mazen Anbari, Wendy S. Lader and William Scott Biel. (Editors), *Child, Parent, and State: Law and Policy Reader*. Philadelphia, PA: Temple University Press, 3-26.
- Chesler, Phyllis. (1986) *Mothers on Trial: The Battle for Children and Custody*. New York: Harcourt Brace Jovanovich.
- Cox, Daniel. (2002) "The Pro Se Puzzle", *LSC's Equal Justice Magazine*, 1, 3.
- Cyprus Mediation Association. (2006). www.cymedas.com/english/family_mediation.php.
- Danforth, Gay and Bobbie L. Welling. (1996). (Editors) *Achieving Equal Justice for Women and Men in the California Courts*. San Francisco: Judicial Council of California Advisory Committee on Gender Bias in the Courts. www.courtinfo.ca.gov/programs/access/documents/f-report.pdf
- Dickey, Anthony. (1997) *Family Law*. Third Edition. Sydney: LBC Information Services.
- Dowd, Nancy E. (2002) "Changing Family Realities, Non-Traditional Families and Rethinking the Core Assumptions of Family Law", in Andrew Bainham (Editor), *The International Survey of Family Law*. 439-469. Bristol, U.K.: Jordan Publishing.
- Farrugia, Ruth. (2001) "It's All Happening in Family Law in Malta", in Andrew Bainham (Editor), *The International Survey of*

- Family Law. (2001 Edition)*, 285-299. Bristol, U.K.: Jordan Publishing.
- Fortin, Jane. (2006) "Accommodating Children's Rights in a Post Human Rights Act Era", *Modern Law Review*, 69, 3, 299-326.
- Funken, Katja. (2002) "Comparative Dispute Management: Court-Connected Mediation in Japan and Germany", *German Law Journal*, 3, 2.
- Glennon, Lisa. (2001) "Family Law: A Process of Reform", in Andrew Bainham ed., *The International Survey of Family Law*, 333-362. Bristol, U.K.: Jordan Publishing.
- Goundry, Sandra; Yvonne Peters and Rosalind Currie. (1998) *Family Mediation in Canada: Implications for Women's Equality. A Review of the Literature and Analysis of Data from Four Publicly Funded Canadian Mediation Programs*. Ottawa: Status of Women Canada.
- Grillo, Trina. (1991) "The Mediation Alternative: Process Dangers for Women", *Yale Law Journal*, 100, 1545-1610.
- Hart, Barbara. (1990) "Gentle Jeopardy: The Further Endangerment of Battered Women and Children in Custody Mediation", *Mediation Quarterly*, 7, 317-330.
- Hassall, Ian and Gabrielle M. Maxwell. (1992) *A Children's Rights Approach to Custody and Access*. Wellington, NJ: Office of the Commissioner for Children.
- Horwill, F.M. and Sophy Bordow. (1983) *The Outcome of Defended Custody Cases in the Family Court of Australia*. Family Court of Australia: Research Report 4.
- Jordan, Peter. (1996) *The Effects of Marital Separation on Men - 10 Years On*. Family Court of Australia: Research Report 14.
- Keilitz, Susan L., Henry W.K. Daley and Roger A. Hanson. (1992) *Multi-State Assessment of Divorce Mediation and Traditional Court Processing*. State Justice Institute, National Center for State Courts.
- Kelly, Joan B. (1989) "Mediated and Adversarial Divorce: Respondents' Perceptions of Their Processes and Outcomes", *Mediation Quarterly*, 24, 71-87.
- Kelly, Joan B. (1990) "Is Mediation Less Expensive? Comparison of Mediated and Adversarial Divorce Processes", *Mediation Quarterly*, 8, 1, 15-26.
- Kuhn, Jeffrey A. (1998) "A Seven-Year Lesson on Unified Family Courts: What We Have Learned Since the 1990 National Family Court Symposium" *Family Law Quarterly*, 32, 1, 67-93.
- Leadership Council on Child Abuse and Interpersonal Violence.
www.leadershipcouncil.org/1/pas/1.html
- Logue, Patricia M. (2002) "The Rights of Lesbian and Gay Parents and Their Children", *Journal of the American Academy of Matrimonial Lawyers*, 18, 95-129.
- Macfarlane, Deborah. (2004) "Family Mediation in France", *Journal of Family Studies*, 10, 97-111.
- Marcus, Mary G., Walter Marcus, Nancy A. Stilwell and Neville Doherty. (1999) "To Mediate or Not to Mediate: Financial Outcomes in Mediated Versus Adversarial Divorces", *Mediation Quarterly*, 17, 143-152.
- Marsiglio, William, Paul Amato, Randal D. Day, and Michael E. Lamb. (2000) "Scholarship on Fatherhood in the 1990s and Beyond", *Journal of Marriage and the Family*, 62, 1173-1191.
- McNeely, Cynthia. (1998) "Lagging Behind the Times: Parenthood, Custody, and Gender Bias in the Family Court", *Florida State University Law Review*, 25, 1-48.
- Minamikata, Satoshi and Teiko Tamaki. (2002) "Family Law in Japan during 2000", Andrew Bainham ed. *The International Survey of Family Law. (2002 Edition)*, 221-228. Bristol, U.K.: Jordan Publishing.

- Moloney, Lawrie. (2000) *Do Fathers 'Win' or Do Mothers 'Lose'? A Preliminary Analysis of a Random Sample of Parenting Judgements in the Family Court of Australia*. Presentation at Australian Institute of Family Studies. (21 September. www.aifs.gov.au/institute/seminars/moloney.html)
- Mosten, Forrest S. (1997) *The Complete Guide to Mediation: The Cutting-Edge Approach to Family Law Practice*. American Bar Association: The Section of Family Law.
- National Center for State Courts. (1997) *Domestic Violence and Child Custody Disputes: A Resource Handbook for Judges and Court Managers*. Williamsburg, VA: National Center for State Courts.
- National Council of Juvenile and Family Court Judges. (1995) *Resource Guidelines: Improving Court Practice in Child Abuse and Neglect Cases*. Reno, NV: National Council of Juvenile and Family Court Judges.
- Nicholson, Alastair and Margaret Harrison. (2000) "Family Law and the Family Court of Australia: Experiences of the Previous 25 years", *Melbourne University Law Review* 756.
- Nygh, Peter. (1985) "Sexual Discrimination in the Family Court", *University of New South Wales Law Journal*, 8, 62-79.
- Pearson, Jessica. (1991) "The Equity of Mediated Divorce Agreements", *Mediation Quarterly*, 9, 179-197.
- Petre, Donna M. (1999) "Unified Family Court: A California Proposal Revisited", *Journal of the Center for Children and the Courts*, 1999, 161-168.
- Rasnow, Tina L. (2002) "Traveling Justice: Providing Court Based Pro Se Assistance to Limited Access Communities." *Fordham Urban Law Journal*, 1281, February.
- Florida Supreme Court. *Report on Gender Bias Study Commission*. (1990). www.flcourts.org/sct/sctdocs/bin/bias.pdf
- Resnik, Judith. (1996) "Asking About Gender in Courts", *Signs*, 21, 952-960.
- Richards, Martin. (1994) "Divorcing Children: Roles for Parents and the State", in Mavis Maclean and Jacek Kurczewski, (Editor), *Families, Politics and the Law: Perspectives for East and West Europe*, 305-320. Oxford: Clarendon Press.
- Sales, Bruce Dennis; Connie J. Beck and Richard K. Haan. (1993) *Self-Representation in Divorce Cases*. Chicago, IL: American Bar Association.
- Senate Legal and Constitutional References Committee. (1998) *Inquiry into the Australian Legal Aid System: Third Report*. Canberra: Parliament of Australia.
- Smart, Carol and Selma Sevenhuijsen. (1989) (Editors) *Child Custody and the Politics of Gender*. London: Routledge.
- Stacey, Judith. (1998) "Gay and Lesbian Families: Queer Like Us", in Mary Ann Mason, Arlene Skolnick, and Stephen D. Sugarman (Editors), *All Our Families: New Policies for a New Century*, 117-143. New York: Oxford University Press.
- Star, Leonie. (1996) *Counsel of Perfection: The Family Court of Australia*. Melbourne: Oxford University Press.
- Wallerstein, Judith S., Julia M. Lewis and Sandra Blakeslee. (2000) *The Unexpected Legacy of Divorce: A 25 Year Landmark Study*. New York: Hyperion.
- Wallerstein, Judith S. and Sandra Blakeslee. (2003) *What About the Kids? Raising Your Children Before, During, and After Divorce*. New York: Hyperion.
- Websites**
- Australian Government. *Families*. www.ag.gov.au/agd/www/familylawhome.nsf
- Canadian Government. *Status of Women*. www.swc-cfc.gc.ca

Conflict Transformation and Peace-Building.

www.peacemakers.ca/bibliography

Custody Disputes. *Keeping Kids Safe.*

www.cavnet2.org/partners1.cfm?partnerid=15263

Family Court of Australia.

www.familycourt.gov.au

Aspasia Tsaoussis

ALBA Graduate Business School

Athens, Greece

atsaoussi@vivodinet.gr

atsaouss@alba.edu.gr

Gender Equity

Irene van Staveren

History

The Role of the UN

Gender has become part of national and international governance debates and regulations since the 1970's. The UN Decade for Women 1976-1985, was initiated in Mexico City. In 1985 and 1995 the UN organized well-attended international conferences on the position of women worldwide in, respectively, Nairobi and Beijing. 1985 is also the year that marks the emergence of a genuinely global women's movement, initiated by women in the South, under the name of DAWN (Development Alternatives with Women for a New era).

The network presented a manifesto, written by Gita Sen and Caren Grown (1985) from the perspective of poor women in developing countries, challenging dominant economic policies and criticizing the debt problem, war, the colonial heritage, food-fuel-water crises, and pleading for women's reproductive rights, the fulfillment of basic needs, and women's empowerment. The pamphlet also challenged North-based women's perspectives on women and development, arguing that the focus should not be on poor women in the South but instead on powerful financial institutions in the North that constrain women's development in the South. This challenge has been taken up by NGOs such as WIDE (Women in Development Europe) and WEDO (Women's Environment and Development Organization).

In the 1990s, women's reproductive rights became a major theme in the women's movement in the South, in particular around the 1994 UN conference on population and development in Cairo. At that conference, there was a strong representation of women's

groups from the global South, who not necessarily defended the same issues as feminists from the North (see, for example, Sonia Corrêa and Rebecca Reichmann, 1994). Another theme that emerged in the women's movement was that of sustainability and its intersections with women's roles in families as the primary care givers.

The attention to sustainability was fuelled by yet another UN conference, on the environment, held in Rio de Janeiro in 1992. A feminist perspective on the environment emerged, which helped the women's movement to shape its views on the complex relationships between women's empowerment, the environment, and caring roles in the context of economic policies oriented towards growth and liberalization (Harcourt 1994; Braidotti et al. 1994). In the words of Wendy Harcourt (1994:7): "In defining a feminist perspective of sustainable development, women can better understand their relationship to the environment as a product of the success and failures of the current play of knowledge/power in the development discourse."

The 1995 UN women's conference brought together more than ten thousand women in Beijing—policy makers, researchers, activists, and representatives of grass-root organizations—probably the largest gathering of the global women's movement in history. It was here, that the UN organization for development, UNDP, presented its annual *Human Development Report* focused on gender, including a gender-aware development indicator to which I shall turn later. In the words of Hilkka Pietilä and Jeanne Vickers (1994:44), "human development" provided the "ultimate focus" for the UN's efforts to improve gender equity worldwide. The outcome of the Beijing conference, often referred to as the *Beijing Platform for Action*, consisted of a long list of strategic objectives and actions and

accompanying institutional and financial arrangements (UN 1996). The areas concerned are poverty, education, health, violence, armed conflict, the economy, decision-making, institutions, rights, the media, and the environment, with special attention to the girl-child.

Since the Beijing conference the emphasis has shifted from negotiating new agreements at the global level towards the implementation of what has been agreed at the major UN conferences of the 1980s and 1990s. This is not to say that there are no more gatherings of women's movements at the international level. To the contrary, the advancement of global communication and transportation over the past decade have stimulated the gathering of women's NGOs, activists, and researchers at meetings across the globe. One such occasion is the Social Forum which is organized regularly at local, national and international level.

In other words, while the official, UN-based efforts have shifted from design to implementation and monitoring, the global women's movement has continued the dialogue—partly to support the implementation of the successes achieved on paper in Rio de Janeiro, Cairo, and Beijing, and partly to discuss new issues which require feminist responses. In the course of such global women's movement gatherings, various alliances were developed between North-based and South-based women's NGOs, such as, for example, the Women's International Coalition for Economic Justice (see, for an overview of women's movements in various regions of the world: Geertje Lycklama à Nijeholt *et al.*, 1998; for a recent study of women and development in relation to the UN, see Devaki Jain 2005).

From 'Women' to 'Gender'

Whereas in the 1970s and 1980s, activists, scholars and policy makers frequently

referred to women and women's disadvantaged position as compared to men, today, the language has shifted to gender equity. The notion of gender equity refers to changes in the social relations between men and women as characterized by power, structures, identities, and meanings. While equality between men and women was the overall objective of the earlier decades, later developments in the women's movement as well as in gender studies have drawn attention to differences between women, multiple identities and the inter-relatedness of social stratifications, such as those of gender, class, and ethnicity. As a consequence, gender equity has emerged as a more open concept than equality, as it goes beyond the political goals of equal access to and control over resources for men and women: it also includes attention to women's agency, in terms of their identities, roles, and participation in decision making, and it emphasizes women's capabilities to choose the lives they have reason to value.

This shift from 'women' to 'gender' parallels the shift in paradigms for the understanding of women's position in society, in particular in the developing world. The dominant paradigm in the 1970s and 1980s was the Women in Development (WID) approach. This approach was based upon two philosophical traditions: on the one hand modernization including the heritage of colonialism, and on the other hand liberalism. The modernist perspective has led to a view of women as lacking behind, as disadvantaged and vulnerable, whereas the liberal view has provided the WID approach with the perspective of integrating women into development through increased labor force participation and an emphasis on individual decision making through markets (Chowdhry 1995). Indeed, critiques of the WID approach precisely point at the hidden assumptions of women being outside the

economy, as non-productive, and the need for them to increase their participation in markets (Mohanty 1991; Hirschman 1995).

Well into the 1990s, the World Bank still held on to the WID approach, for example through its policy paper *Enhancing Women's Participation in Economic Development* (World Bank, 1994). From the critiques on the WID approach a new approach emerged under the label of Gender and Development (GAD), which emphasizes women's empowerment in a variety of domains of life, including the economy but not limited to women's economic position. Moreover, the GAD approach pays much more attention to the role of power and meaning in the relations between men and women—in the household, culture and the media, social relations, institutions, and policy making. It is this approach which remains relevant today for understanding gender issues, for recognizing intersections of gender with class and ethnicity, and for informing gender equity policies.

The shift from WID to GAD was partly informed by gender analyses of macroeconomic policies in the 1980s and first half of the 1990s, in particular Structural Adjustment Policies (SAPs) advised by IMF and World Bank to developing countries with debt crises and macroeconomic instability. Numerous studies of SAPs and their intended and unintended effects on poor men and women's lives have made clear that simply increasing women's participation in the labor market—in wage labor or self-employment—did not help much to reduce women's poverty. In fact, SAPs have reinforced gender inequalities by shifting public investments away from the government budget to women's unpaid work and by liberalization policies that made women workers end up at the bottom of increasingly flexible global production systems (Bakker 1994; Sparr 1994; Elson

1995; Standing, 1999). The social safety nets which SAPs provided for disadvantaged groups appeared to be far from sufficient to reduce women's poverty. Moreover, the liberalization policies that formed part and parcel of SAPs have been identified as a major reason for the Asian financial crisis of 1997 (Stiglitz 2002).

This crisis led to a disproportionate incidence of unemployment among female workers, as well as an increase in women's unpaid workload (UNDP 1999; UNIFEM 2000; World Bank 1998). For example, right after the onset of the Asian crisis in 1997 the Korean government urged women to 'Get Your Husband Energized' (Singh and Zimmit 2000:1260), even though 86 per cent of job losses in the banking and finance sector were women's jobs (World Bank 1998). Therefore, Diane Elson and Nilüfer Çağatay have remarked, "Creditors were in effect 'bailed out' while poor women acted as unpaid provisioners of last resort" (2000:1355).

The history of the global governance of gender equity was clearly written in the global South, which hosted the historical international conferences and led the paradigm shift from WID to GAD, but which also suffered most from the negative gender impacts of structural adjustment and financial crises.

Gender Equity Goals and Measurement

UN Indicators

In 1995, The UN introduced a gender-disaggregated version of its widely used Human Development Index (HDI): the Gender related Development Index (GDI). This indicator is an index with a value between zero and one, which incorporates gender differences in the three variables that make up the HDI: education (school enrolment and literacy rates), health (life expectancy) and income (GDP per capita).

Each of the variables is measured separately for men and women, whereby the income variable is based on estimates of women's labor force participation and the average gender wage gap per country. The higher the inequality, the lower the GDI ranking compared to a country's HDI rank. Like the HDI, the GDI is published annually by UNDP in its *Human Development Reports*, with a country ranking that signals the extent of gender inequality in human development. See Table 1, below:

Table 1. HDI & GDI Rankings, Selected Countries

Country	HDI	GDI	HDI-GDI
Poland	43	22	+21
Thailand	48	33	+15
Jamaica	66	52	+14
Spain	8	34	-26
Saudi Arabia	61	81	-20
Netherlands	4	20	-16

Source: Adapted from Dijkstra and Hanmer (2000)

As can be seen from Table 1 the HDI and GDI rankings differ, but not substantially – the ranking is quite similar because of the fact that a major component of both is the level of per capita income. In order to redress this weakness, Geske Dijkstra and Lucia Hanmer (2000) have proposed an alternative measure. This is the Relative Status of Women indicator (RSW) which uses exactly the same variables as the HDI and GDI, but only the gender differences and not their absolute values. This results in a significantly different ranking of countries: poor countries with good records of gender equality now can score higher than rich countries with substantial inequalities. This is particularly striking for countries that have relatively low female labor force participation (such as the Netherlands) or a relatively high gender wage gap (such as Japan): these countries have

lower RSW rankings than much poorer nations. See Table 2, below:

Table 2. GDI & RSW Rankings, Selected Countries

Country	GDI	RSW	GDI-RSW
Vietnam	91	13	+78
Tanzania	111	37	+74
Lithuania	57	4	+53
United Arab Emirates	39	99	-60
Greece	22	75	-53
Costa Rica	31	76	-45

Source: Adapted from Dijkstra and Hanmer (2000)

Another gender indicator introduced by the UNDP helps to measure women's status: the Gender Empowerment Measure (GEM). This index consists of three dimensions of empowerment: economic participation and decision making, political participation, and decision-making and power over economic resources. The index consist of four indicators: (1) share of seats in parliament held by women; (2) female share of legislators, officials and managers; (3) share of female professional and technical workers; (4) ratio of estimated female to male income. The GEM serves as an indicator of women's status independent from the income level of a country. For 2006, the ten countries with the highest scores for women's empowerment are all Northern European countries, with relatively good scores for some developing countries, such as Trinidad and Tobago and Namibia. Among the low GEM scores are not only low-income countries, but also Japan, Turkey, the Russian Federation, Saudi Arabia, South Korea and Malta (UNDP 2006).

Finally, since the 1995 women's conference, UN agencies have put much effort in collecting data for a whole range of other gender-disaggregated variables. The

OECD has set up an online database with about fifty gender indicators, the Gender Institutions and Development database (GID). Indeed, various, though not all, indicators in the database refer to institutions such as family law, property rights, or harmful practices such as female genital mutilation. The data are categorized into four categories: family code, physical integrity, civil liberties, and ownership rights. Information on cultural and traditional practices that impact upon women's economic development has been translated into an index in order to measure the level of discrimination and allow comparison between countries.

Women's Capabilities

The indicators discussed above—developed and published by the UN on an annual basis at country level, and through national *Human Development Reports* also for regions and social groups within countries – are useful for aggregate level gender analysis and monitoring of gender equity policies. But micro level research requires a more refined and contextual set of goals and measures. It is in the work of Amartya Sen on the Capability Approach that gender equity finds a prominent place in the analysis of poverty and development (Agarwal, Humphries and Robeyns 2006). Sen's work stresses that wellbeing should not be equalized with income, nor with utility maximization or basic needs, but with the capabilities that people have. Such a focus brings to the fore the need to measure poverty at the intra-household level, to take differences into account between women's agency and their achieved functionings, and points at the important role of various institutional constraints and opportunities that may be gender-biased.

Whereas Sen does not engage in listing particular capabilities, Martha Nussbaum (2000) has developed a list of ten capabilities

from a gender-aware perspective. Her list consists of the following capabilities: (1) life (2) bodily health (3) bodily integrity (4) senses, imagination, and thought (5) emotions (6) practical reason (perception of the good and critical reflection about the planning of one's life) (7) affiliation (to others and from others to oneself) (8) other species (9) play (10) control over one's environment (political and material).

Nussbaum recognizes that for the realization of equal capabilities for everyone, some rule is necessary about priorities. She finds such rule in John Rawls' (1971) maximin criterion of fairness. This criterion states that inequality can only be allowed when the activities driving the inequality benefit the most disadvantaged. Applying this criterion to the Capability Approach, Nussbaum proposes a minimum threshold for each capability, linked to country's constitutions. Policies for furthering capabilities should therefore prioritize to get everyone across the threshold level for each capability, before spending resources on further increases of capabilities that are likely to benefit some, but not all.

Whereas various gender researchers have found the Capability Approach helpful in analyzing and evaluating differences in the wellbeing of women and men, others find Nussbaum's work—in particular her list—too universalistic. In a rich empirical study about capabilities of women in the United Kingdom, Ingrid Robeyns (2003) has followed Sen's approach of finding out people's valued capabilities through discussions. In that study, Nussbaum's list was only partially confirmed. See Table 3, below, for examples of specific measures of capabilities and functionings for men and women in the UK. The table shows gender differences for physical as well as social capabilities and functionings.

Table 3. Selection of Capabilities and Functionings for Women and Men of Different Ages, UK, 1998

Capability/ Functioning	Men 15- 65 years	Women 15-65 years	Men >65 years	Women >65 years
Hearing problems	6.8%	3.6%	30.5 %	21.2%
Diabetes	1.8%	1.7%	8.3%	5.9%
Never go out after dark	-	-	17.6 %	50.1%
Can borrow money from someone	72.9 %	77.3%	57.5 %	57.3%

Source: Adapted from Ingrid Robeyns (2002)

Although the disparity between the capabilities listed by Nussbaum, informed by interviews with women in India, and the ones found in discussions with women's groups in the UK by Robeyns is not very large, there are a few significant differences between Nussbaum's list and Robeyns' findings, in particular relating to the value of time and childcare. The approach, however, is still developing and its usefulness for gender equity policy is yet to be established.

Gender Equity through Mainstreaming

It is the GAD approach that gave way to today's policy efforts of gender mainstreaming, as the genuine integration of gender equity in a wide variety of policy areas. Such mainstreaming is not about integrating women in existing policies, institutions and structures, but is driven by an "agenda-setting" vision, as Rounaq Jahan (1995) has formulated it, implying "the transformation of the existing development agenda with a gender perspective" (Jahan 1995:13). In other words, today's approach towards gender equity entails that "women not only become part of the mainstream, they also reorient the nature of the mainstream" (ibid).

Gender mainstreaming implies a holistic view of how gender impacts on society and

how, at the same time, social, cultural, political and economic processes have impacts upon the relationships between women and men. Hence, gender equity is understood in a layered way, not just as being about equal opportunities or equal outcomes. Gender implies, first, a shaping of market processes in terms of men's and women's differential access to and control over resources. Second, gender involves the shaping of people's choices, for example in segmented labor markets, through which men 'choose' typically masculine jobs and women 'opt for' typically feminine jobs. Third, gender is recognized as part of macroeconomic trends, for example through fluctuations in the female labor force participation rate or in responses to crises though increases in the supply of unpaid labor.

In short, in international political economy, gender is understood not only as an exogenous variable (coming from outside the economic system, from culture, social relations, nature, or laws), but increasingly also as an endogenous force – shaping and being shaped by particular economic processes, conditions, and outcomes.

In the early 1990s gender mainstreaming efforts were mainly undertaken in social areas of policy making, such as health care, education, and social welfare programs, not economic areas such as finance, transport or trade. The social policy areas were, at least for a long time, the primary responsibility of the state, and were for a large part governed through the public sector, financed by tax revenues and user fees. Later in the 1990s, and particularly in the new millennium, gender mainstreaming has reflected the increasing consciousness among policy makers that also other policy areas have potential gender dimensions. Moreover, the continued trend of globalization of capital, production, labor, services, and even utilities,

has increased the awareness that today almost every area of life is affected by a reduction in national government provisioning and regulation. This openness to global changes—for better or worse—is not gender-neutral, but carries gender biases, which makes gender mainstreaming even more important than it was before the current era of globalization.

Mainstreaming Gender Equity in Trade

Gender analysis of trade provides a good illustration of how gender plays out at the global level and what the challenges are for improving gender equity in international trading relations (van Staveren, Elson, Grown, and Çağatay, 2007). A remarkable trend in globalization is that the global workforce has become notably female. Women make up about 50 percent of migrant workers, mostly as nurses, nannies, and domestic workers. At the same time, the majority of workers in labor intensive export industries in developing countries is female (Joekes & Weston 1994). In developed countries export employment has declined in labor-intensive sectors which have moved to developing countries due to the wide availability of low skilled female labor and low women's wages there. Women in OECD countries have particularly lost jobs in female-intensive sectors such as textiles, garments and leather products (Kucera & Milberg 2000).

The reasons for the predominantly female employment in export industries in developing countries relate to perceived qualities of female labor as well as to more explicit forms of labor market discrimination. Women tend to be regarded by employers as having 'nimble fingers' and being more obedient than men and therefore as more suitable for work in typical low-skilled, low paid, repetitive jobs in export manufacturing. In this sector, women's wages are

considerably lower than men's wages. The gender wage gap provides employers with a cost advantage, without much loss of productivity as long as women are hired only for low-skill jobs. Econometric analysis has shown that countries with the highest ratio of export earnings to their GDP have also the largest gender wage gaps, even when educational levels of women have moved close to those of men (Seguino 2000).

Women are not only the preferred workforce in wage employment for exports, but also make up the majority of flexible workers—own account workers, home workers, sub-contractors, day-laborers and seasonal workers—in the global production system. Women find themselves increasingly working informally at the bottom of global value chains, at low labor standards, outside the reach of labor laws, labor inspectors, and company codes of conduct (Barrientos et al 2003).

The expansion of trade may also have impacts on unpaid domestic work, which is primarily done by women and girls. Tax revenue from trade taxes has fallen in many developing countries which has led to reductions in public expenditure on infrastructure, education and health (Khattry 2003). In turn, this is likely to increase women's unpaid work burden, as women will try to produce substitutes for reduced public services in order to keep up household living standards (Benería & Feldman 1992; Bakker 1994; van Staveren 2002).

Mainstreaming gender equity in trade requires elimination of wage discrimination, investment in women's education, as well as wider opportunities for women to upgrade their small scale businesses or to move to skilled jobs in wage employment, but gender mainstreaming also requires (re)investment in the care economy. The existing mainstreaming of gender in international labor conventions, such as those of the ILO

(International Labor Organization) are not sufficient to ensure gender equity in the world of work in the near future. Additional efforts need to be supported at the international level, including company codes of conduct, the formulation of gender equity goals and measures in trade agreements, and a re-shifting of resources from trade to domestic sectors of the economy, including support for unpaid caring.

Mainstreaming Gender Equity in Poverty Reduction Strategies

Around the turn of the millennium, Poverty Reduction Strategy Papers (PRSPs) appeared on the stage for highly indebted countries and other countries seeking loans from IMF and the World Bank. This follow-up of SAPs provided a good opportunity for mainstreaming gender into macroeconomic policies in developing and transition economies. Many PRSPs have recognized gender as a relevant dimension of poverty. But at the same time gender assessments of PRSPs are rather negative about the effectiveness of the attention to gender in many PRSPs. Ann Whitehead (2003) has examined how gender issues have been taken into account in the PRSPs of Tanzania, Bolivia, Malawi and Yemen. Her conclusion is simply that “gender is not integrated, or mainstreamed” (Whitehead 2003:35).

In an evaluation of thirteen PRSPs, Elaine Zuckerman and Ashley Garrett conclude that “it is encouraging to see progress in mainstreaming gender into PRSPs ... but it is still far from adequate” (Zuckerman and Garrett, 2003:12). Moreover, they continue, “No PRSPs try to assess the gender implications of structural adjustment measures such as state-owned enterprise privatization and trade liberalization”(ibid:7). In addition, the evaluation studies reveal that most PRSPs, contrary to the advice in the

PRSP *Sourcebook*, have no chapter on gender.

The PRSP *Sourcebook* (Klugman 2002) explains that the heart of the PRSP should entail a macroeconomic framework that promotes low inflation, a sustainable balance of payment position, growth, and a sustainable fiscal position. In chapter 10, the *Sourcebook* provides PRS teams with relevant, up-to-date, and extensive information for the integration of gender (Bamberger, Blackden et al 2001). But, despite the comprehensive gender advice available to PRSP teams, gender remains largely absent in PRSPs.

The gender issues which *are* present in PRSPs, are only found, to some extent, in the participatory process, the poverty analysis, and to a limited extent in the proposed social policies, while completely absent in the macroeconomic framework of the PRSP – the heart of the PRSP. It therefore seems that it is the narrow macroeconomic framework that constrains genuine integration of gender in the PRSP. This macroeconomic framework is not a neutral set of policies but is embedded precisely in the wider, neoliberal policy environment supported by the Washington Consensus (van Staveren 2008).

This not only implies a missed opportunity for achieving gender equity at the macroeconomic level, but it also ignores that poverty reduction and economic development partly depend on the elimination of gender inequalities and the empowerment of women. For example, growth has been affected negatively by gender inequality in education (Klasen 1998). Currency devaluations tend to stimulate women’s employment in labor intensive export industries, while at the same time a devaluation will make imported food more expensive, which puts pressure on women’s role as household food providers in many countries around the world.

Deflationary policies are not gender-neutral either, as they tend to have negative feedback effects on women's wage employment, survival of small businesses depending on small expensive loans, and support from public services. Finally, contractionary policies aimed at reducing budget deficits are likely to hurt those groups in society that are most dependent upon redistributive policies through public expenditures, including women, given their gender role as carers (Elson and Çağatay, 2000). Moreover, women already tend to be disadvantaged by gender biases in public expenditures, as gender audits of government budgets have shown (Norton and Elson, 2002). Hence, budget cuts tend to re-inforce the male bias in public expenditures.

Mainstreaming Gender Equity in Public Finance

Various countries across the world have experience with gender analyses of public finance, such as Australia, South Africa, Tanzania, and England (Budlender, 1996; Sharp and Broomhill, 1990; TNGP, 1999). The starting point of these gender analyses of public finance were studies of state expenditures and their impact on women, and assessments of government employment policy in terms of equal opportunity. Later gender budget studies included also an analysis of tax revenues. The gender analysis of public finance allows for an assessment of the (allocative) inefficiency inherent in government budgets and caused by gender discrimination. By asking questions, about the direct and indirect impacts and the equity and efficiency outcomes of government budgets on women and men, gender analysis forces re-evaluation of a long-held assumption that government budgets and economic policies generally are 'gender neutral' (Sharp 1999).

Diane Elson (1997) has summarized the requirements which allow gender budget analyses (GBA) to become effective tools within the political process:

1. Take into account the effect of a special investment or public spending in general along gender lines, including a closer look to which extent men and women make different use of the policies offered.

2. Measure the effect of public spending in each sector on male and female well-being. In other words, what is the impact on income, livelihood, nutrition, and human capital investment?

3. Include a gender disaggregated tax incidence analysis, both of income tax and VAT attached to the different goods and services provided by the state.

4. Concentrate also on non-monetary dimensions of people's well-being, particularly time use. What is the effect on the total productive time of men and women? Here, not only productive work in return for monetary income should be taken into account but also unpaid labor, since women tend to work more hours than men when paid and unpaid hours are added.

5. Take into account feedback effects on the economy as a whole. What are the effects of public expenditures reflecting a specific development strategy, such as the shift from import-substitution to export promotion strategy? Or what are the gender impacts of a reduction of the welfare state on labor supply, productivity, human and social capital formation?

The governance implications of a gender analysis of public finance, or GBA, are not only of a distributional character. It is important to emphasize that GBAs also may re-inforce a positive relationship between equity and efficiency. On the distributional side, non-discrimination in social spending will help to reduce poverty more effectively than spending patterns that benefit the already

advantaged groups. For example, in education, financing gender equality in primary schooling is likely to be more effective than the funding of an increase in university students. On the allocative efficiency side, which is just as important for the reduction of women's poverty, gender analyses of public finance point out that equal access to resources will prevent under-investments in human and financial capital, while equal access to services, such as electricity and agricultural extension services, will increase the marginal labor productivity of women's paid and unpaid work (Krug & van Staveren 2002).

Millennium Development Goal on Gender Equity

In the year 2000, the UN led the Millennium Development Goal (MDG) initiative that found support from almost every country in the world as well as from leading global institutions, including World Bank and the IMF. The MDGs form a set of eight goals on poverty, education, health, and the environment, for the year 2015. Goal number three is concerned with gender equity, that is, to promote gender equality and empower women. The specific target for MDG 3 is, however, much more limited, namely to eliminate gender disparity in primary and secondary education, preferably by 2005, and in all levels of education no later than 2015. The monitoring indicators, however, include the share of women in wage employment outside agriculture, and the share of seats by women in parliament.

The relevance of MDG 3 for the promotion of gender equity is firstly, that the goal, as part of all MDGs, has been widely agreed, and secondly, that there is close monitoring of the targets so that policy makers have continuous information about the feasibility of the goal for each country and a possible need to increase efforts. Critics,

however, argue that the complexity of gender equity cannot be reduced to just one goal. In addition, a critique on the MDGs in general is that they seem to ignore how Washington Consensus policies undermine the very basis of the MDGs. For example, privatization of schools and student funding is likely to benefit middle class families, to the disadvantage of children from poor households, in particular girl children whose brothers are often given priority when family sources for advanced schooling are limited. In the words of Elson (2004:7) "Nevertheless, though the Millennium Development Compact goes beyond the Washington Consensus in advocating a bigger role for public investment and more debt relief, it does not call into question the benefits of liberalization of international trade and finance. A limitation of the approach of the Millennium Project Secretariat is an overemphasis on a set of 'interventions' with insufficient consideration of systemic issues, at both national and global levels".

The UN Taskforce Gender Equality has provided advice on the necessary strategy for the achievement of MDG 3 (Grown and Rao Gupta, 2005). The Taskforce has emphasized that meeting MDG 3 depends on meeting the other goals, whereas the success of the other goals also depends on the extent to which MDG 3 will be met. The Taskforce has identified several strategic priorities for action:

1. Strengthening opportunities for post-primary education for girls, while simultaneously meeting commitments to universal primary education.
2. Increasing adolescents and women's access to a broad range of sexual and reproductive health information and services.
3. Investing in infrastructure designed to reduce women's time burdens.
4. Guaranteeing women's property and inheritance rights.

5. Eliminating inequality in employment by decreasing women's reliance on informal employment, closing gender gaps in earnings, and reducing occupational segregation.

6. Increasing women's shares of seats in national parliaments and local government bodies.

7. Significantly reducing violence against women and girls.

The above strategies are much more diverse than the goal of MDG 3; they are applicable not only to developing countries but also to developed countries; and they largely follow the multiple legacies of the Beijing women's conference of 1995. The Taskforce explicitly links gender equity to efficiency, arguing that the two are not necessarily opposite but may reinforce each other. If, however, MDG3 will not be realized, development will be confronted with considerable costs for the period 2005-2015, as estimated by Klasen and Abu-Ghaida (2004): GDP growth rates will be 0.4 percentage points lower; fertility decline will be 0.6 children per woman less; higher child mortality of 20 to 32 children per 1000; and 20.6% more underweight children.

Finally, the Taskforce recommends that accountability and monitoring of countries' progress on MDG 3 should involve the global women's movement as stakeholders, and institutionalize monitoring through the well-established UN mechanism for country reports and high-level meetings on gender equity, namely the Convention on the Elimination of Discrimination Against Women (CEDAW).

Conclusion

Gender equity has become part and parcel of development policy through the efforts of gender mainstreaming since the 1990s. Today, every government has some form of gender policy through which it has translated the recommendations of the 1995 women's

conference. International institutions have also taken responsibility for the *Beijing Platform for Action*, although the level of mainstreaming is often limited to the social sectors. The global governance gap in gender equity now lies with the economic sectors, such as industrial policy, trade policy, and international finance. Today, there is an urgent need for a next stage of gender mainstreaming. This stage requires much international coordination, in order to reverse the trend of a global and highly gendered race to the bottom, driven by the gender wage gap, flexibilization of work, food insecurity, and an overburdened care economy.

Selected References

- Agarwal, Bina; Jane Humphries and Ingrid Robeyns. (2006) (Editors) *Capabilities, Freedom and Equality: Amartya Sen's Work from a Gender Perspective..* Delhi/Oxford: Oxford University Press.
- Bakker, Isabella. (1994) (Editor) *The Strategic Silence. Gender and Economic Policy*. London: Zed Books.
- Bamberger, M.; M. Blackden; L. Fort and V. Manoukian. (2001) "Gender" and "Annex I. Gender Technical Notes", in J. Klugman (Editor), *A Sourcebook for Poverty Reduction Strategies*. Washington D.C.: World Bank.
- Barrientos, Stephanie; Catherine Dolan and Anne Tallontire. (2003) "Gendered Value Chain Approach to Codes of Conduct in African Horticulture", *World-Development*, Volume 31, Number 9, pp. 1511-26.
- Beneria, Lourdes and S. Feltman (1992) (Editors) *Unequal Burden: Economic Crises, Persistent Poverty and Women's Work*. Boulder: Westview Press.
- Braidotti, Rosi; Ewa Charkiewicz; Sabine Häusler and Saskia Wieringa. (1994) *Women, the Environment and Sustainable Development. Towards a Theoretical*

- Synthesis*. London/Santo Domingo: Zed/INSTRAW.
- Budlender, Debbie. (1996) (Editor) *The Women's Budget*. Cape Town: Institute for Democracy in South Africa.
- Chowdhry, Geeta. (1995) "Engendering Development? Women in Development (WID) in International Development Regimes", in M. Marchand and J. Parpart (Editors), *Feminism, Postmodernism, Development*. London: Routledge, 26-41.
- Corrêa, Sonia and Rebecca Reichman. (1994) *Population and Reproductive Rights. Feminist Perspectives from the South*. London/New Delhi: Zed/Kali for Women.
- Dijkstra, Geske and Lucia Hanmer. (2000) "Measuring Socioeconomic Gender Inequality: Toward an Alternative to the UNDP Gender-Related Development Index", *Feminist Economics*, Volume 6, Number 2, pp. 41-75.
- Elson, Diane. (1995) (Editor) *Male Bias in the Development Process*. Manchester: University of Manchester Press.
- Elson, Diane. (1997) *Integrating Gender Issues into Public Expenditure: Six Tools*. Paper by the Genecon Unit, University of Manchester. Graduate School of Social Sciences.
- Elson, Diane. (2004) *The Millennium Development Goals. A Feminist Development Economics Perspective*. The Hague: Institute of Social Studies, 52th Dies Natalis Address.
- Elson, Diane and Nilüfer Çağatay. (2000) "The Social Context of Macroeconomic Policies", *World Development*, Volume 28, Number 7, pp. 1347-64.
- Grown, Caren and Geeta Rao Gupta (2005) *Taking Action: Achieving Gender Equality and Empowering Women*. London: Earthscan.
- Harcourt, Wendy. (1994) "Introduction", in Wendy Harcourt (Editor), *Feminist Perspectives on Sustainable Development*. London/Rome: Zed/SID, pp. 1-8.
- Hirschman, Mitu. (1995) "Women and Development: a Critique" in M. Marchand and J. Parpart (Editors), *Feminism, Postmodernism, Development*. London: Routledge, pp. 42-55.
- Jahan, Rounaq. (1995) *The Elusive Agenda. Mainstreaming Women in Development*. Dhaka/London: University Press/Zed.
- Jain, Devaki. (2005) *Women, Development and the UN. A Sixty Year Quest for Equality and Justice*. Bloomington: Indiana University Press.
- Joekes, S. and A. Weston. (1994) *Women and the New Trade Agenda*. New York: UNIFEM.
- Khattry, B. (2003) "Trade Liberalization and the Fiscal Squeeze: Implications for Public Investment", *Development and Change*, Volume 34, Number 3, pp. 401-24.
- Klasen, Stephan. (1998). *Gender Inequality and Growth in Sub-Saharan Africa: Some Preliminary Findings*. Background Paper prepared for the 1998 SPA Status Report on Poverty in SSA. Washington DC: World Bank.
- Klasen, Stephan and Dina Abu-Ghaida. (2004) "The Costs of Missing the Millennium Development Goal on Gender Equality", *World Development*, Volume 23, Number 7, pp. 1075-1107.
- Klugman, J. (2002) (Editor) *A PRSP Sourcebook for Poverty Reduction Strategies*. Washington D.C.: World Bank.
- Krug, Barbara and Irene van Staveren. (2002) "Gender Audit: Whim or Voice?", *Public Finance and Management*, Volume 2, Number 2, pp. 190-217.
- Kucera, David and William Milberg. (2002) "Gender Segregation and Gender Bias in Manufacturing Trade Expansion: Revisiting the "Wood Asymmetry", *World Development*, Volume 28, Number 7, pp. 1191-1210.

- Lycklama à Nijeholt; Geertje; Virginia Vargas and Saskia Wieringa. (1998) *Women's Movements and Public Policy in Europe, Latin America, and the Caribbean*. New York: Garland.
- Mohanty, Chandra Talpade. (1991) "Under Western Eyes: Feminist Scholarship and Colonial Discourses", in C. Mohanty, A. Russo and L. Torres (Editors), *Third World Women and the Politics of Feminism*. Bloomington: Indiana University Press, pp. 51-80.
- Nussbaum, Martha. (2000) *Women and Human Development: The Capabilities Approach*. Cambridge: Cambridge University Press.
- Norton, A. and D. Elson. (2002) *What's Behind the Budget? Politics, Rights and Accountability in The Budget Process*. London: Overseas Development Institute. 54pp.
- Pietilä, Hilka and Jeanne Vickers. (1994) *Making Women Matter. The Role of the United Nations*. London: Zed Press.
- Rawls, John. (1971) *A Theory of Justice*. Cambridge: Harvard University Press.
- Robeyns, Ingrid. (2002) *Gender Inequality. A Capability Perspective*. Cambridge: Cambridge, UK: Cambridge University.
- Robeyns, Ingrid. (2003) "Sen's Capability Approach and Gender Inequality: Selecting Relevant Capabilities", *Feminist Economics*, Volume 9, Numbers 2-3, pp. 61-92.
- Seguino, Stephanie. (2000) "Gender Inequality and Economic Growth: A Cross-Country Analysis", *World Development*, Volume 28, Number 7, pp. 1211-30.
- Sen, Gita and Caren Grown. (1985) *Development, Crises, and Alternative Visions: Third World Women's Perspectives*. Olden: Mediaredaksjonen.
- Sharp, Rhonda and Ray Broomhill. (1990) "Women and Government Budgets", *Australian Journal of Social Issues*, Volume 23, Number 1, pp. 2-14.
- Singh, Ajit and Ann Zammit. (2000) "International Capital Flows: Identifying the Gender Dimension", *World Development*, Volume 28, Number 7, pp. 1249-68.
- Sparr, Pamela. (1994) (Editor) *Mortgaging Women's Lives. Feminist Critiques of Structural Adjustment*. London: Zed Books.
- Staveren, Irene van. (2002) "Global Finance and Gender" in Jan-Aart Scholte and Albrecht Schnabel (Editors), *Civil Society and Global Finance*. London: Routledge, pp. 228-46.
- Staveren, Irene van. (2008) "The Gender Bias of the Poverty Reduction Strategy Framework", *Review of International Political Economy*, Volume 15, Number 2, pp. 287-313.
- Staveren, Irene van; Diane Elson; Caren Grown and Nilüfer Çağatay. (2007) (Editors) *The Feminist Economics of Trade*. London: Routledge.
- Standing, Guy. (1999) "Global Feminization Through Flexible Labor: A Theme Revisited", *World Development*, Volume 27, Number 3, pp. 583-602.
- Stiglitz, Joseph. (2002) *Globalization and its Discontents*. New York: Norton.
- TNGP. (1999) *Budgeting with a Gender Perspective*. Dar es Salaam: Tanzania Gender Networking Programme.
- UN. (United Nations) (1996) *Platform for Action and the Beijing Declaration*. Fourth Conference on Women, Beijing, China, 4-15 September 1995. New York: United Nations.
- UNDP. (1999) *Human Development Report 1999*. Oxford: Oxford University Press.
- UNDP. (2006) *Human Development Report 2006*. Oxford: Oxford University Press.
- UNIFEM. (2000) *Progress of the World's Women 2000*. New York: UNIFEM.

- Whitehead, Ann. (2003) *Failing Women, Sustaining Poverty: Gender in Poverty Reduction Strategy Papers*. London: Gender and Development Network.
- World Bank. (1994) *Enhancing Women's Participation in Economic Development*. Washington D.C.: World Bank. World Bank Policy Paper.
- World Bank. (1998) *East Asia: The Road to Recovery*. Washington D.C.: World Bank.
- Zuckerman, Elaine and Ashley Garrett. (2003) *Do Poverty Reduction Strategy Papers (PRSPs) Address Gender? A Gender Audit of 2002 PRSPs*. Washington D.C.: Gender Action.

Irene van Staveren
Institute of Social Studies, The Hague;
Radbouod University, Nijmegen;
The Netherlands
i.staveren@chello.nl

Geography and Governance

Bin Zhou

Introduction

At the World Bank website focused on best practice in dealing with the social impacts of oil and gas operations, one article states that "Fundamentally, stagnant or diminished development is an issue of *governance*." This is so due to the fact that the state of development is a measure of a society's inability or unwillingness to manage its resources (such as energy) for the purpose of fostering constructive economic, political, and social development (World Bank 2005). The United Nations Development Programme or UNDP defines governance as the exercise of political, economic, and administrative authority in the management of a country's affairs at all levels and views governance and development as indivisible (UNDP 1997). This is so because the purpose of governance is human development and the improvement of the human condition (Streeten 2005).

Governance is not merely how the government manages and regulates the development of a society. The realms of governance include the role of the state, the private sector, and civil society (UNDP 1997). That is, the state of governance reflects the efforts of the entire society in using its resources for certain objectives. Thus, the state of development and human conditions of society represent the state of governance as a result of collective and aggregate efforts of the society. Governance is an effort to improve human conditions on the part of society, individuals, and organizations. On the other hand, human conditions are a manifestation of governance or the consequences of human efforts designed to improve their living circumstances and development prospects.

Human efforts, human resources, and human institutions are always deployed in a certain physical environment. Aspects of this physical environment include location, climate, topography, available water bodies, biomass and biodiversity, mineral resources, and distance from other economic centers. The physical conditions act as the initial context of human development or the external environment underlying civilization. For human production processes that intimately interact with elements of the physical environment, the physical environment inevitably exerts effects on production processes. This raises the question of the role of governance on the one hand, and geography (here mainly physical geography) on the other, in shaping human conditions and affecting the level of development. As Acemoglu et al (2003) state, here the geography refers to the "forces of nature" while institutions or governance is about "man-made" influences. This article addresses this issue from several aspects. It first discusses the role of geography and governance as competing forces shaping development. Then the article discusses the effect of geography and governance as co-existent factors that enable society to develop. This is followed by a discussion of geography as a factor that influences human conditions through governance and institutions. Finally, the article discusses the relationship between governance/institutions and natural disasters as a particular geographical condition.

Geography or Governance?

Although very few people attribute human conditions exclusively to geography or governance, historically and contemporarily there are divergent views regarding which factor plays a primary role in shaping development. The most famous of these may be the environmental determinism view.

Environmental Determinism

Environmental determinism is a view of cultural ecology, of how human-environmental relationships are structured, that sees the physical environment as causing social development (Rubenstein 2005). According to this view, people and peoples are what they are because they have been shaped by their physical surroundings--climate, vegetation, and so on. The notion is considered to be implicitly racist since it associates the superiority of certain peoples from northern latitudes or a cool environment, and denigrates peoples from the tropics.

Environmental determinism has its direct roots in nineteenth century German geography with the work of Friedrich Ratzel (1844-1904) (Ratzel 1882,1891). However, relevant developments associated with its genesis can be traced back to social changes within earlier centuries that were intimately linked with changing geographic ideas (Dickinson 1969; Martin and James 1993).

The sixteenth century marked the beginning of an era that is often characterized as the Age of Discovery. Freed from the Islamic threat to Europe, and fueled by a zeal to spread Christianity, exploration was planned and supported both by European governments and merchant companies (Martin and James 1993). With the twin goals of spreading Christianity and garnering personal wealth and resources, Europeans began to sail into uncharted waters and expand the horizons of their world. This activity had several major impacts on geography and society. Exploration fueled the need for accurate maps and for reliable navigational instruments. Thus, it is in this era that necessity spurred developments in technology. Exploration also resulted in a deluge of new resources and new information. The new resources and personal wealth that resulted from these explorations

ultimately provided a foundation for the growth and restructuring of economy, society, and governments (Dickinson 1969; Martin and James 1993). The geographers of the era fell to the task of recording and compiling the flood of information with great energy, and used it to improve available maps and charts, as well as to compile encyclopedic works. These geographers would be characterized more accurately as cosmographers. They saw no disciplinary boundaries but embraced all knowledge of the natural sciences as legitimate within their compendiums. It was during this era of discovery that European geographers recognized the diversity of both the human and physical world, and began to question the relationships between human activity and the physical environment. It was at the intersection of this questioning, and the rise of modern science towards the end of the nineteenth century, that environment determinism has its genesis (Holt-Jensen 1999).

The two German geographers Alexander von Humboldt (1769-1859) and Carl Ritter (1779-1859) are often said to represent both the end of the classical era of geography and the beginning of the modern (Martin and James 1993). Both Humboldt and Ritter wrote monumental works of synthesis in the classical tradition (Ritter 1822-59; von Humboldt 1845-62), but both also laid the foundations for scientific study within geography by introducing the systematic treatment of data into their work (Dickinson 1969; Holt-Jensen 1999; Martin and James 1993). It was the scientific geography begun by Humboldt and Ritter that Ratzel built upon, and from which environmental determinism sprang.

‘Science’ and ‘scientific geography’ were based on empirical data, analysis, the search for regularities, and ultimately the identification of natural laws. Ratzel was influenced by these ideas and looked for

cause and effect in the relationships between human beings and their environment. In the first volume of his great work *Antropogeographie* he attempted to bring the study of human geography into the scientific era by stressing the causal effect that the environment has on human activity. While Ratzel did modify his environmental determinism in later work, his views were used selectively and distorted for political and ideological ends (Martin and James 1993). In addition, the English ‘translation’ of his *Antropogeographie* by Ellen Churchill Semple (1863-1932) (Semple 1911) emphasized and modified the environmentalist viewpoint, and it was this view that came to dominate American geography until the 1930s (James and Jones 1954; Dickinson 1969; Holt-Jensen 1999).

William Morris Davis (1850-1934), perhaps the most influential American geographer of his era, espoused environmental determinist views but as a geomorphologist his influence in this regard came primarily through his stature and through his involvement in geographic education. It was through the work of Semple, Ellsworth Huntington (1876-1947), and Griffith Taylor (1880-1963)—three human geographers—that environmental determinism became pervasive in American Geography. Both Semple and Huntington were prolific writers and presenters of their ideas. Semple spread her ideas through her university teaching and written works. Her writing had literary flare and was widely read, and she was an equally persuasive teacher, influencing the many students of geography in her classes. Ellsworth Huntington, who was a student of Davis, developed the notion that climate had a causal effect on human development. Civilization could only emerge under stimulating, mid-latitude climatic conditions and by the same logic the heat of tropical regions would prevent higher levels

of civilization from developing. Huntington’s works were widely read outside of geography, and his ideas used in American schools. Taylor’s environmental determinist views were politically unwelcome in Australia where interest in settling the Outback was high. He moved to the United States, and later to Canada. It was in America that he found a fertile ground to develop his determinist ideas. He insisted that his work was built on scientific knowledge, and remained committed to the determinist view long after most geographers rejected it (Holt-Jensen 1999; Martin and James 1993).

Revival of Environmental Determinism?

After the early 20th century, environmental determinism has been largely discredited due to its use of loose correlations and anecdotal evidence, the tendency to ignore contrary evidence, and the largely ethnocentric ranking of the environment. However, toward the end of the 20th century, largely due to developments external to geography, the role of geography in economic development has received rising attention. Some development and growth economists, in their efforts to model divergent economic growth patterns across the world, attribute significant roles to differences among countries in terms of geographical conditions. For example, particular geographical circumstances—whether a country is landlocked and not open to trade—will permanently inhibit markets, limit economies of scale, lower its efficiency and ultimately growth and development (Sachs and Warner 1995a,b, 1997). Natural resources such as minerals and ecological conditions favoring cash crops also affect income (Easterly and Levine 2003). Bloom and Sachs (1998) point to Africa’s tropical location as largely a hindrance to development. Bloom and Sachs (1998) and Sachs (2001) argue that tropical location leads to underdevelopment mainly

due to (1) soils that are fragile and of lower fertility; (2) the prevalence of crop pests and parasites; (3) excessive plant respiration and a lower rate of photosynthesis; (4) high evaporation rates and insufficient supply of water; (5) lack of a dry season, lack of cold temperatures, or insufficient length of summer days for temperate crop growth; (6) ecological conditions favoring diseases that infect humans and livestock; (7) lack of coal deposits; and (8) high transport costs. Landes (1998) emphasizes the inhibiting effects of high temperatures on humans' willingness to work. These are collectively called the "Geography/Endowment Hypothesis" (Coviello 2003).

The most sweeping view regarding the dominant role of the physical environment since the late 20th century is found in Jared Diamond's influential book *Guns, Germs, and Steel*. Diamond recognizes the proximate causes of Western dominance since the 16th century as emanating from guns, germs, and modern technologies (steel, etc.) that Westerners used to decimate indigenous societies in the New World and Africa. Diamond looks for deeper causes that led to the rise of these proximate causes since the last Ice Age 13,000 years ago, in his search for an ultimate explanation of the broadest patterns of history. Diamond zeros in on the availability of wild ancestors of plants and large mammals, their suitability for domestication, on different continents, and the orientation of continents.

For example, in the Fertile Crescent, wild ancestors of many crops were abundant, highly productive, and occurred in large stands. Many species of cereals and pulses developed as annuals, putting much energy into producing big seeds. Domestication was relatively easy, with little additional change needed for cultivation. Many of these seeds were edible by humans, which presented their value clearly to hunters and gathers. Cereals

and pulses domesticated in the Fertile Crescent accounted for 6 of the modern world's 12 major crops. These advantages made the big seeded cereals the first crops developed in the Fertile Crescent. As Diamond points out, the favorable local flora was not an isolated factor in securing the Fertile Crescent's food production system. Other factors, such as climate, environment, and animals were all important ingredients. Blumler (1992) studies the world distribution of large seeded grass species. These are wild grasses with significant seed size that might potentially be chosen as targets for domestication by human ancestors. Of the 56 species he found, 32 are native to the Mediterranean zone or other seasonally dry environments, especially the Fertile Crescent or other parts of western Eurasia's Mediterranean zone. In contrast, East Asia has only 6 species, and Sub-Saharan Africa has 4. The Americas have 11 species with North America having 4, Mesoamerica 5, and South America 2.

The availability of wild plants that are easy to domesticate, along with some necessary accompanying conditions for plant domestication, is at the heart of Diamond's explanation of history's broadest pattern. Similarly, western Eurasia had a high concentration of ancestors of wild large mammal species for domestication. It had 72 such species, compared with 51 in Sub-Saharan Africa, 24 in the Americas, and 1 in Australia. Of all 14 domesticated large mammals, 13 were confined to Eurasia, becoming an indispensable component of the food production system. The east-west orientation of the Eurasian continent provided ease in the spread of this productive agricultural system from the Fertile Crescent to Europe, North Africa, and the Indus Valley.

A productive food production system was only the first the step in a long sequence of

development. According to the Diamond's logic, the developed food production system was also responsible for high population densities. Epidemic diseases evolved from the dense population of domestic animals with which humans came into close contact. The immunities humans in the Old World acquired allowed them to survive while the germs could pass on to and kill the indigenous peoples of the New World upon close contact with Old World invaders. In addition, high population densities also triggered the development of political organization, which helped effective territorial governance and political-military expansion. Developed agriculture also enhanced the division of labor, promoted invention, innovation, and technological development, which eventually gave rise to guns, cannons, and other weapons.

Thus, a conquering and killing machine formed based on the early advantages in the rise of food production. Such a mechanism essentially explains why it is the West that dominates the other world regions instead of some other region. Diamond's message has been further reinforced by the biogeographical evidence of Hibbs and Olsson (2003) and Olsson and Hibbs (2005) who found that different initial conditions in biogeography and geography largely account for the different timing of the Neolithic transition and thus ultimately help account for the large divergent income levels among nations today. They also found that the effect of geography is only partly mediated by the quality and performance of institutions.

Primacy of Governance

Institutionalists hold that institutions are the first-order determinants of economic performance (Coviello 2003), or as Rodrik et al call it, the institutional primacy over geography. Although not using the word "governance", the institutional elements

contained in a typical model are in large measure those associated with governance. For example, Kaufmann et al (1999a, 1999b) and Easterly and Levine (2003) design an institutional index that contains the effect of six institutional (governance) measures: public voice and accountability; political stability and absence of violence; government effectiveness; extent of regulatory burden; rule of law; and freedom from graft and corruption. They also adopt three indicators of the performance of the macroeconomic environment: openness, real exchange rate overvaluation and inflation. Other measures they adopt include ethnolinguistic diversity and religion.

Institutionalist views can also claim a long tradition tracing back to John Locke, Adam Smith, and John Stuart Mill (Acemoglu et al 2006). For example, John Locke stressed the importance of property rights, and emphasized the role of government in "preservation of the property of ...member of society" (Locke (1690) 1980:47). Economists Gunnar Myrdal (1974) and Douglas North (1993) were each awarded the Noble Prize in part for articulating the role of institutions. In their study, Easterly and Levine (2003) find strong evidence to support the primacy of institutions or governance over geography in shaping economic performance. Their study does not support the idea that tropical location and lack of access to the sea inhibit development. Rather, the institutional index significantly explains economic development, consistent with the institutions hypothesis. In addition, in their study, the strong positive impact of institutional development on economic development is also robust in altering instrumental variables in the model. Coviello (2003) also found evidence consistent with the institutionalist view rather than the geography/endowment hypothesis.

Rodrik (2002) and Rodrik and Submaranian (2003) estimate the respective

contributions from geography, institutions, and trade in determining income levels around the world. Their results show that the quality of institutions "trumps" everything else. Once institutions are controlled for, measures of geography have only very weak impacts on income. These results lead the authors to declare "Institutions rule" as reflected in the view of the primacy of institutions over geography in economic development.

Geography and Governance – Possibilism

While previous sections discuss views that see geography and governance as two competing forces shaping economic development and human conditions, some scholars tend to emphasize both. A traditional geographical view of cultural ecology embodies the well-known Possibilist perspective.

Possibilism views the relationship between human beings and the environment as one of reciprocity. While the environment certainly both constrains and enables human activity, the environment is not causally related to human action. It is possible for people to choose from many courses of action and even to alter their environment. The possibilist view is attributed to the French school of geography, in particular to the French historian Lucien Febvre (1878-1956). However, it is Paul Vidal de la Blache (1845-1918), who dominated French geography for over forty years, who is largely responsible for the establishment of possibilism. Vidal de la Blache contended that while the environment sets limits and offers possibilities for human action, humans also react and adjust to given environmental conditions in various ways, depending on their 'genres de vie' – i.e., culture; the complex learned traditions, institutions, technologies, attitudes, and values. This variety of possibilism came to challenge

environmental determinism especially through the work of Isaiah Bowman (1878-1950) and Carl Ortwin Sauer (1889-1975) (Barton and Karan 1992; Holt-Jensen 1999; Martin and James 1993).

While Bowman began his career as a disciple of William Morris Davis, by 1919 he had repudiated determinism. He became widely published, and is perhaps best known for his work in political geography. He was frequently called upon by the U.S. government for his advice. Amongst many other political activities he served as Chief Territorial Advisor to the American Commission to Negotiate Peace after World War I, and served as Special Advisor to the Secretary of State in World War II. Within these political forums he had ample opportunity to argue the possibilist view and oppose determinism (Barton and Karan 1992; Holt-Jensen 1999; Martin and James 1993).

Carl Sauer became influential through his leadership of the Berkeley or Western School of American geography. Sauer is best known for his work on the morphology of the landscape, and for his acknowledgement of the importance of history and cultural context of human settlement. He espoused the principle that the same environment could have very different meanings to people depending on their attitudes, objectives, and level of technology. Current landscapes could be seen to be the result not just of nature, but also of repeated and varying cultural impressions over time. For Sauer, any separation of environment and human activity was flawed since they exist in a recursive relationship that varies both temporally and spatially (Barton and Karan 1992; Holt-Jensen 1999; Martin and James 1993).

At the beginning of the twenty-first century, geographers are once more contemplating issues central to cultural ecology. The rising interest in the study of political ecology has pointed to the dialectical

link between physical environmental factors and institutions/governance. Researchers look at the society and culture not only in their natural environment, but also in their political environment as well. They also look at how unequal relations among societies, and within a society, affect the natural environment (Billon 2001; Toly 2004). Environmental determinism may no longer be rejected outright, or possibilism accepted without question. Contemplating the relationship between human beings and environment is once more the subject of vigorous debate.

Challenges to "Institutions Rule"

Even some economists reject views expressed in the primacy of institutions perspective. Naude (2004) estimates the effects of geography, policy, and institutions in the context of African economic development. He found that while institutional factors such as literacy, investment, foreign direct investment, government expenditure, and urban agglomeration clearly have significant impacts on economic development, geographic factors such as settler mortality as impacted by tropical ecology, the degree of 'landlock', land area, and malaria also exert significant roles. The results reject the notion of "either institutions or geography"; that one is more important than the other in the context of economic development, and contradicts, at least for Africa, the finding of Rodrik et al. (2004) that "institutions rule." (2004: 842).

Sachs, who is generally viewed as a geography/endowment hypothesist (Easterly and Levine 2003), at times also expresses views that support a more balanced approach toward the role of geography and institutions in economic development. He offers a broad explanation of divergent levels of economic performance among countries, focusing on three major categories of factors. The first category is geographical factors. He points

out that certain parts of the world are geographically favored with advantages ranging from natural resources, coastlines, navigable rivers, proximity to economically advanced nations, and advantageous conditions for agriculture and human health. However, Sachs equally emphasizes the role of socio-cultural systems and the cumulative effects of history in shaping development levels (Sachs 2000).

The other category of factors is that of social system (governance or institutions). He states that certain social systems have supported modern economic growth, whereas others have not. Precapitalist systems based on selfdom, slavery and inalienable landholding have tended to slow modern economic growth. Centralized state 'socialism' (e.g., the USSR) during the 20th century eventually came to inhibit economic well-being and growth. Colonial rule was generally adverse to high rates of economic growth.

In addition to these two groups of factors, Sachs also points to the role of history in reinforcing previous development and exerting a cumulative feedback factor, in line with recent "New Economic Geography." Positive feedback processes amplify the advantage of early industrialization and widen the gap between rich and poor, by allowing the early developers to conquer and exploit the late developer, causing some late developers to collapse. The technological gap between early developers and late developers also tends to widen over time. In this sense, the historical or feedback factor discussed by Sachs can be broadly placed within the institutions or governance view, since it involves the issue of how to govern the political and economic relationships among today's nation states, that at some point in history were colonies and colonizers.

Although Sachs (2003) recognizes that institutions matter, he warns that the

characteristics of institutions do not explain everything. He is critical of those studies that attribute most development problems to institutions, at the expense of geography and resource constraints. He sees this as oversimplifying the issue of development by resorting to a single factor explanation. He warns of the danger of an institutional argument that would relieve rich countries of financial responsibility for the poor, because development failures are seen to be the result of institutional failures rather than lack of resources (Sachs 2003:38). Sachs emphasized the need to combat AIDS, tuberculosis, malaria, soil depletion and building more roads to connect remote populations to regional markets and coastal ports. He criticizes some economists (Acemoglu *et al.* 2001) who see malaria as a minor factor in Sub-Saharan Africa because most adults have some acquired immunity. This is seen as neglecting the negative impact of the disease on local investment, transaction costs, trade and tourism. In general, constraints due to the physical environment and geographical isolation are still a large part of the developmental problem in places such as Sub-Saharan Africa and the Andean countries of Latin America (Gallup 2000). Sachs calls for development thinking recognizing both institutions (including governance) and resource endowment as critical (Sachs 2003:41).

Geography through Governance

Geography and governance have more subtle relationships, instead of simply competition or coexistence. Specifically, geography may impact on institutions and governance. A classic example of how geography may effect governance is the Nazi interpretation of environmental determinism. After the unification of Germany in 1871, Ratzel produced a major work on political geography (Ratzel 1897) arguing that states

evolve in the manner of organisms out of evolutionary necessity. Ratzel's ideas were embraced by those who saw them as providing a scientific rational for the German right to control neighboring states and to view nations and races as existing in a natural hierarchy. These geopolitical notions ultimately became important within Nazi ideology and action (Holt-Jensen 1999). Thus, in late nineteenth and early twentieth century geography, politics, and government became inextricably linked.

It also has been noted that embracing an environmental determinist philosophy was more than merely an intellectual exercise. Environmentalist ideas served to retroactively rationalize expansionist socio-political objectives and lessen American guilt on the treatment of African Americans. They also served to legitimize U.S. imperialistic goals. Thus, environmental determinism had social and geopolitical causes and consequences in both old and new worlds. It was not until the ideological extremes and associated atrocities of the Nazis became apparent that environmental determinism was rejected. In recent decades it has become acceptable to explore strong connections between environment and human activity.

In the early years of the new millennium the role of institutions and environment on development are both recognized. Rodrik *et al.* (2004), Rodrik (2002) and Rodrik and Subramanian (2003) adopt a scheme to analyse the relationship between income, geography, and institutional factors. They consider endowment and productivity as endogenous factors, trade and institutions as partly endogenous factors, and geography as an exogenous factor. The reason institutions are partly endogenous is that even institutionalists recognize the role of geography in building institutions.

Institutionalists contend that geography and environment impact on economic

development through long-lasting institutions. For example, environments where crops are most effectively produced using large plantations, will quickly develop political and legal institutions that protect large landholders from the many peasants and small landowners, and may even develop slavery as a way to accommodate the needs of the large landholders (Engerman and Sokoloff 1997; and Sokoloff and Egerman 2000). In places like this, even when agriculture is no longer the main economic activity, enduring institutions may continue to inhibit competition and economic development for the majority of the population.

Similarly, the era of colonization laid the foundation for many of the institutions of today, which provides a natural experiment for geography to shape institutions (Acemoglu et al. 2001, 2002, 2006). For example, colonizers established institutions where a few settlers manipulated a local political power structure in order to exploit the material riches. Especially where extractive industries dominated, democratic legal institutions were inhibited (e.g., sub-Saharan Africa). In contrast, in places where settler institutions were established (e.g., US, Australia), the rule of law, property rights, and democracy were eventually established. Thus, a process of institutional reversal was in place where the originally more developed became less developed while originally less developed regions became more prosperous (Acemoglu et al. 2001, 2002, 2006).

The institutional structures created by colonialists in response to environment thus endure with the end of colonialism. Institutionalists argue the impact of the environment on economic development runs through its long-lasting institutions. Most institutionalists do not completely deny the role of geography and the environment on economic development. However, they reject

a claim for the direct role of geography. Instead, they trace the role of geography and environment through institutions (Easterly and Levine 2003).

Many researchers have found evidence of the role of geography in affecting institutions. For example, Hall and Jones (1999) associate the lack of development in tropical countries with the fact that Europeans did not settle in the tropics and thus did not bring high quality institutions to these areas. Here the point is that it is not the tropical environment per se, but the lack of high quality institutions, that hinders development. Acemoglu et al (2001) suggest that European settlement in North America, Australia, and New Zealand created institutions to support private property rights and limit the power of the State. On the other hand, Europeans did not settle in the Congo, Burundi, the Ivory Coast, Ghana, Bolivia, Mexico, Peru, etc. They established institutions that empowered an elite to extract gold, silver, cash crops, etc. For Easterly and Levine (2003), slavery was a way for Europeans to capture a labor force for extractive states, such as in the Caribbean and Brazil. Acemoglu et al (2006) notice that Pilgrims settled in the American colonies instead of Guyana partially because of the high mortality rates there. Similarly, Sokoloff and Egerman (2000) find that a Puritan colony on Providence Island off the coast of Nicaragua did not last long. Sokoloff and Egerman (1997) and Egerman and Sokoloff (2000) provide evidence to support a crop hypothesis. They argue that land endowments in Latin America lent themselves to commodities favoring economies of scale, and/or the use of slave or local indigenous labor (sugar cane, rice, silver). They are thus historically associated with power concentrated in the hands of the plantation and mining elite.

In contrast, land endowments in North America lent themselves to commodities

grown on family farms (wheat and maize), and thus promoted the growth of a large middle class which helped spread power widely. Once the power structure was formed, the elite in Latin America created institutions that preserved their hegemony, such as restricting voting rights, public land and mineral rights distribution, and limiting access to schooling. Even granting new corporate charters favor those with elite insiders. These elite groups ultimately were opposed to democracy. In contrast, North America enjoyed a larger middle class with a less powerful elite, so that the United States and Canada created more open and egalitarian institutions featuring broader voting rights, equal protection before the law, wider distribution of public lands and mineral rights, lower entry barriers to businesses, and a big government effort to provide schooling. Differences in institutions between Latin America and North America also contributed to the larger European immigration flows to North America than to Latin America (Sokoloff and Egerman 2000).

Natural Disasters and Institutions

Previous discussions focus mostly on the long term interaction and relationship between geographical conditions and governance/institutions. In the short run, natural disasters happen, as particular forms of geographical phenomena or conditions. Natural disasters include hurricane, flood, thunderstorm, tornado, storm surge, landslide, mudslide, earthquakes, tsunamis, cold, heat, avalanche, disease, drought, winter storm, waterspout, famines, fire, hail, volcanic eruption, etc. These are calamities or catastrophes that may result in a disruption of the normal functioning of social life, and cause property damages and even the loss of human lives. Some 75% of people in the world live in areas affected at least once by earthquake, tropical cyclones, flood or drought between

1980 and 2000. Billions of people in more than 100 countries are periodically exposed to at least one event of earthquake, tropical cyclone, flood or drought and more than 180 deaths per day occurred as a result of these disasters (UNDP 2005). 160 countries have more than a quarter of their population living in areas at relatively high mortality risk from one or more natural hazards. More than 90 countries have over 10% of their population residing in areas at relatively high risk from at least two natural hazards. In 35 countries at least 5% of the population is at a relatively high risk from three or more natural hazards (Hotspots 2004). Of all natural disasters, flood risk was found to affect world's greatest land area (14.4 million sq km²), population (3.9 million), amount of GDP (\$22,859 billion), amount of agricultural GDP (\$528 billion), and length of roads and rail (1.5 million km). Asset at risk from other types of natural disasters are one or two orders of magnitude smaller (UNDP 2005; ISDR 2004). From 1985 to 2003, about one third of the world's land area including 82% of the world population was exposed to flood. The most flood prone areas encompass 9% of land area and more than 38% of world's population. The most flood prone area was Asia including countries such as the Philippines, Indonesia, Vietnam, Thailand, Cambodia, the Korean Peninsula, China, India, and Bangladesh. Countries in other areas include Eastern Africa, Europe, coastal South America, Central America and the U.S. Midwest (ISDR 2004).

The 1990s witnesses a worsening trend of human suffering and economic losses from natural disasters. The total number of people each year affected by natural disasters nearly doubled between 1990 and 1999, by an average of 188 million per year (CRED 2002). Asia is disproportionately affected by natural hazards with approximately 43% of all natural disasters, and with 70% of natural

disaster related deaths, in the decade from 1991 to 1999. During the two El Nino years of 1991-92, and 1997-98, floods in China alone affected over 200 million people in each year. In 2004, a tsunami, originated in the southwest coast of Sumatra Island, sent tidal waves across the Indian Ocean affecting countries such as Indonesia, Thailand, India, Sri Lanka, etc. The death toll is over 200,000 people. Between 1990 and 1999, natural disasters had multiplied fourfold but economic losses were 14 times higher. Most economic losses occurred in developed countries, but seen as losses by percentage of GDP, developing countries lost most in relative terms.

Ironically, though natural hazards appear to be naturally occurring events, humans, guided by their cultural institutions, are always a factor that causes these natural events “disasters.” Natural disasters are determined as much or more by societal behavior and practice as by nature per se (SNDR 1996). UNDP (2005) summarizes the interactions between natural disaster and social/economic development into three categories: disaster limits development (destruction of fixed assets, health and education infrastructure, and personnel), development causes disaster risk (unsustainable developments that cause degradation of the environment; and development paths that promote social separation), and development reduces disaster risk (increasing access to wider resources or wider social groups reduces the vulnerability). Indeed, in 1982, Hurricane Isaac destroyed 22% of the housing stock in the Tongan archipelago, setting back the development effort of the island. New Orleans was under water for at least a month in 2005 following the storm surge from Hurricane Katrina that burst the levee. The inaction of the federal and state government to the local newspaper’s prediction of pending danger from possible hurricane

attacks plays a significant role. With increasing levels of technology, additional resources, and improved coordination among all stakeholders, the disaster risk reduction will be enhanced (ISDR 2002). Humans, follow their cultural institutions, live in and adapt to natural environment. They will inevitably encounter natural disasters.

Although natural disasters do not choose who the victims are, human institutions tend to arrange the living space in a way that puts members of lower well being in more dangerous situation than those of higher well being. For example, from 1980 to 2002, while 11% of people exposed to earthquake, tropical cyclones, flood, or drought lived in countries classified by the UNDP as lower human development, these countries accounted for 53% of total deaths to these disasters (ISDR 2004). ISDR also finds that flood hazard risk tend to associate with countries having high proportion of their populations in low income and low population density. A systematic examination of mortality as a result of cyclones, drought, earthquake, flood, landslide, and volcanic eruption in Africa, East Asia and Pacific, Europe and Central Asia, Latin America and the Caribbean, Middle East and Africa, North America, and South Asia reviews that lower and lower middle suffer higher mortalities per 100,000 persons from 1981 to 2000 (UNDP 2005). ISDR (2002) finds that between 1975 and 2001, 90% of natural disaster related deaths are to be found in developing countries.

Institutional factors such as level of literacy and education, the existence of peace and security, access to basic human rights, systems of good governance, social equity, positive traditional values, knowledge structures, customs and ideological beliefs, and overall collective organizational systems, economic status of individuals, communities, and nations all contribute to vulnerability to

natural disasters, along with physical and ecological factors (ISDR 2002).

The rising natural disasters have shaped the agenda of the international community, national governments, and local communities. The United Nations declared the 1990s the International Decade for Natural Disaster Reduction and advanced work to reduce the consequences of natural disaster, under the theme *Building a Culture of Prevention*. The *World Conference on Natural Disaster Reduction*, Yokohama, convened in 1994 and passed resolution the *Yokohama Strategy and Plan of Action for a Safer World* stressing that every country had the sovereign and primary responsibility to protect its people, infrastructure, and national social and economic assets from the impact of natural disaster. The initial emphasis in the natural disaster reduction was placed on the role of science and technology in monitoring the natural hazards, and developing an understanding of their continuing changing patterns, and developing tools and methodologies for disaster reduction. However, over time, the limitations of science and technology in response to the problems of people and political processes in identifying and managing risk factors have become clear. An increasing global awareness is to engage a broader community in hazard awareness and risk management, and wider participation by local communities in hazard and risk reduction activities.

This new awareness has given rise to a new institutional framework toward disaster reduction. The institutional framework incorporates elements of policy, legislation, and organizational development at the national and local decision-making. Such a framework aims at encouraging widespread decision-making and participation, in addition to the pivotal role of government. National, state or provincial governments and legislation are seen as leading the policy

direction and laying the legal foundation. Regional cooperation is important for places having similar geographical conditions to share their resources and combine their experiences in disaster risk reduction. However, local based decision-making processes should also be designed and related to the larger administrative processes so that the actual professional and human resources are delivered on the ground. Local communities should have their own plan of action in disaster risk reduction. Local leadership should integrate the local planning into the large administrative and resource capability at the state or provincial level, mobilize local residents, non-profit organizations and volunteers in enhancing risk awareness, and building local based collaboration, self-reliance and mutual assistance. A comprehensive survey by the United Nations reveals that many countries have established a certain form of national center for natural disaster monitoring and management. However, during the Hurricane Katrina in 2005 in the United States, there were a general breakdown in communication among different levels of government, a broad lack of sufficient actions from local, state to federal government in monitoring and responding to the disaster, and a significant magnitude of misuse of resources. This illustrates that there is still a long way to go for individual countries to work out a particular natural disaster risk reduction institutional framework within their national context.

Conclusion

The intersection of geography and governance is an area that still requires further research. Geographers should be encouraged to break out of the fear of environmental determinism to explore the geography of governance as it is affected by local environment and cultural practices. For

economists, the current focus on cross-section data modeling may be too limited. As some have hinted (Gullap 2000; Naude 2004) the role of geography may differ with different levels of development and the nature of national economies. For countries in early development stages and with a large income share coming from land-based resources, geography may indeed play a larger role. This may explain the broad pattern of development Diamond (1997) and Olsson (2005) have observed for prehistorical and historical times. However, for societies increasingly depend on investment, innovation, human capital, and technological development for the improvement of human conditions, geography and the physical environment may no longer provide first order causes for economic performance. Instead, institutions and governance that promote investment and innovation may indeed play a larger role.

Selected References

- Acemoglu, D.; S. Johnson and J.A. Robinson. (2001) "The Colonial Origins of Comparative Development", *American Economic Review*, 91, 1369-1401.
- Acemoglu, D.; S. Johnson and J.A. Robinson. (2002) "Reversal of Fortune: Geography and Institutions in the Making of the Modern World Income Distribution", *Quarterly Journal of Economics*, 117, 1231-1294.
- Acemoglu, D.; S. Johnson and J.A. Robinson. (2006) "Understanding Prosperity and Poverty: Geography, Institutions, and Reversal of Fortune", in Abhijit V. Banerjee, Roland Benabou, and Dilip Mookherjee,, (Editors), *Understanding Poverty*. New York: Oxford University Press, 19-36,
- Barton, Thomas F. and P.P. Karan. (1992) *Leaders in American Geography*. Mesilla, NM: New Mexico Geographical Society.
- Billon, P.L. (2001) "Political Ecology of War: Natural Resources and Armed Conflicts", *Political Geography*, 20, 561-584.
- Bloom, D.E. and J.D. Sachs. (1998) "Geography, Democracy, and Economic Growth in Africa", *Brookings Papers on Economic Activity*, 2, 207-273.
- Blumler, M. (1992) *Seed Weight and Environment in Mediterranean-Type Grasslands in California and Israel*. UMI Dissertation Services Ann Arbor, MI.
- Coviello, D. (2003) *Instrumental Variables Regressions in Growth, Geography, and Institutions: Reconsidering Some Results*. The World Bank Conference of Sarajevo. October.
- Diamond, Jared. (1997) *Guns, Germs, and Steel: The Fates of Human Societies*. New York: W.W. Norton.
- Dickinson, Robert E. (1969) *The Makers of Modern Geography*. New York: Frederick A. Praeger.
- Easterly, W. and R. Levine. (2003) "Tropics, Germs, and Crops: How Endowments Affect Economic Development", *Journal of Monetary Economics*, 20, 3-39.
- Engerman, S. and K. Sokoloff. (1997) "Factor Endowment, Institutions, and Differential Paths of Growth among New World Economies", in S.H. Harbor (Editor), *How Latin America Fell Behind?* Stanford, CA: University Press, 260-304.
- Gallup, J.L. (2000) "Geography and Socioeconomic Development", *The Andean Competitiveness Project*. Harvard University: Cambridge, MA.
- Hall, R.E., and C.L. Jones. (1999) "Why Do Some Countries Produce so much Output per Worker than Others?", *Quarterly Journal of Economics*, 114, 83-116.
- Hibbs, D.A. Jr. and O. Olsson. (2003) "Geography, Biogeography, and Why Some Countries Rich and Others Poor?", *Economic Sciences*, 101, 3715-3720.

- Holt-Jensen, Arild. (1999). *Geography. History and Concepts. A Student's Guide*. Fourth edition. London: Paul Chapman Publishing Ltd.
- Humboldt, Alexander von. (1845-62) *Kosmos: Entwurf Einer Physischen Weltbeschreibung*. 5 vols, Stuttgart; Tübingen: Cotta.
- International Strategic Disaster Reduction. (ISDR) (2004) *Vision of Risk: a Review of International Indicators of Disaster Risk and its Management*. New York: United Nations.
- International Strategic Disaster Reduction. (ISDR) (2002) *Living with Risk: a Global Review of Disaster Reduction Initiatives*. New York: United Nations.
- James, Preston E. and Clarence F. Jones (1954) *American Geography. Inventory and Prospect*. Syracuse University Press, Syracuse.
- Kaufmann, D.; A. Kraay and P. Zoido-Lobaton. (1999a) Aggregating Governance Indicators. World Bank Research Working Paper 2195.
- Kaufmann, D.; A. Kraay and P. Zoido-Lobaton. (1999b) *Governance Matters*. World Bank Research Working Paper 2196. Washington DC; World Bank.
- Landes, D. (1998) *The Wealth and Poverty of Nations*. New York: W.W. Norton.
- Locke, J. [1690] 1980. *Two Treatises of Government*. Indianapolis: Hackett.
- Martin, Geoffrey J. and Preston E. James (1993) *All Possible World. A history of Geographical Ideas*. Third Edition. New York: John Wiley & Sons.
- Naude, W.A. (2004) "The Effects of Policy, Institutions and Geography on Economic Growth in Africa: An Econometric Study of Based on Cross-Section and Panel Data", *Journal of International Development*, 16, 821-849.
- Olsson, O. and D.A. Hibbs Jr. (2005) "Biogeography and Long-run Economic Development" *European Economic Review*, 49, 909-938.
- Ratzel, Friedrich. (1882) *Antropogeographie: I, Oder Grundzüge der Anwendung der Erdkunde auf die Geschichte*. Stuttgart: Engelhorn.
- Ratzel, Friedrich. (1891) *Antropogeographie: II, Die Geographische Verbreitung des Menschen*. Muenchen: Oldenbourg.
- Ratzel, Friedrich. (1897) *Politische Geographie*. Murnchen: Oldenbourg.
- Ritter, Carl. (1822-59) *Die Erdkunde im Verhaeltniss zur Natur und zur Geschichte des Menschen: oder allgemeine vergleichende Geographie, als sichere Grundlage des Studiums und Unterrichts in physikalischen und historischen Wissenschaften*, 21 vols. Berlin: G. Reimer.
- Rodrik, D. (2002) (Editor) *Searching for Growth: Analytical Narratives of Growth*. Princeton NJ: Princeton University Press.
- Rodrik, D. and A. Subramanian. (2003) "The Primacy of Institutions (and what this does and does not mean)", *Finance & Development*, 40, 31-34.
- Rodrik, D.; A. Subramanian and F. Trebbi. (2004) "Institutions Rule: The Primacy of Institutions over Geography and Integration in Economic Development", *Journal of Economic Growth*, 9, 131-165.
- Sokoloff, K.L. and S.L. Engerman. (2000) "Institutions, Factor Endowment, and Differential Paths of Development in the New World", *Journal of Economic Perspectives*, 14, 217-232.
- Rubenstein, James M. (2005) *An Introduction to Human Geography. The Cultural Landscape*. Fifth Edition. Upper Saddle River, NJ: Pearson Prentice Hall.
- Sachs, J. (2003) "Institutions Matter, But Not for Everything", *Finance & Development*, 40, 38-41.
- Sachs, Jeffery. (2000) "Notes on A New Sociology of Economic Development", in

- Lawrence E. Harrison and Samuel P. Huntington (Editors), *Culture Matters: How Values Shape Human Progress*. . New York: Basic Books.
- Sachs, J. (2001) *Tropical Underdevelopment*, National Bureau of Economic Development Working Paper 8119. New York: NBER.
- Sachs, J. and A. Warner. (1995a) "Economic Reform and the Process of Global Integration", *Brookings Papers on Economic Activity*, 1, 1-95.
- Sachs, J. and A. Warner. (1995b) *Natural Resource Abundance and Economic Growth*. National Bureau of Economic Research Working Paper 5398. New York: NBER.
- Sachs, J. and A. Warner. (1997) "Fundamental Sources of Long-Run Growth", *American Economic Review: Papers and Proceedings*, 87, 184-188.
- Semple, Ellen C. 1911. *Influences of Geographical Environment*. Henry Holt, New York.
- Streeten, P. (2005) *Global Governance for Human Development*. Occasional Paper 4. New York: UNDP.
- Subcommittee on Natural Disaster Reduction. (SNDR) (1996) *Natural Disaster Reduction: a Plan for the Future*. Washington, D.C., National Science and Technology Council.
- Toly, N.J. (2004) "Globalization and Capitalization of Nature: Political Ecology of Biodiversity in Mesoamerica", *Bulletin of Science, Technology & Society*, 24, 47-54.
- United Nations Development Programme. (UNDP) (1997) *Governance for Sustainable Human Development*. New York: UNDP.
- United Nations Development Programme (UNDP) (2005) *Natural Disaster Hotspots: a Global Risk Analysis*. New York: United Nations.
- World Bank (2005) *Governance and Human Rights: an Overview*. New York: World Bank.

Bin Zhou
 Department of Geography
 Southern Illinois University
 Edwardsville, Illinois, USA
 bzhou@siue.edu

Global Sex Sector

Alys Willman-Navarro

Introduction

Around the world, the commercial sex industry is a booming business, providing employment for millions of people and generating multi-billion dollar profits. With globalization, the industry has expanded and diversified, further entangling itself with other economic sectors and involving an ever-wider circle of actors. Yet policymakers generally remain reluctant to treat the sex industry as an integrated social and economic sector, preferring instead to address it indirectly through health policy or law enforcement. This artificial separation often results in inconsistent and even contradictory policies that fail to address key governance challenges ranging from human rights violations, to gender discrimination and employment policies.

The term “sex sector” was coined by a 1998 survey by the International Labor Office of four Southeast Asian countries (Lim 1998). The term refers to an integrated economic sector centered on commercial sex that is connected to economic and social life and that contributes, directly or indirectly, to economic growth, employment and national income. Because the sex sector is deeply rooted in social relations, particularly gender norms, the organization of the sector varies from place to place. The use of the term sector here allows for the inclusion not only of the sex *industry* - defined by Agustin (2005:681) as “all commercial goods and services of an erotic and sexual kind” - but also the surrounding relationships that connect commercial sex to economic, political and social life.

Understood in this broad sense, the sex sector involves a wide range of actors, services and institutions. Products and

services include prostitution; print, video and Internet pornography; stripping and erotic dance; erotic massage; fetish services; live sex shows and cybersex acts; bondage and domination services; organized international sex tours; and escort and call girl services. Participants in the sex sector include not only those who directly sell sex, but also those who benefit directly and indirectly from it. These include, for instance, clients; intermediaries such as business owners and managers, pimps, madams and taxi drivers; newspapers, magazines and Internet sites (which profit from advertising sex-related businesses); telephone and cable companies; hotel chains; travel agents and tourist operators; police and public officials who may accept bribes for turning a blind eye to illegal activities or solicit sexual services; waitresses, bouncers and bartenders in establishments where sex is sold; and families and partners of sex workers who depend on earnings from commercial sex. The majority of commercial sex workers everywhere are women, but the presence of male (Padilla 2007; Kaye 2007) and transgender workers (Kulick 1998) is growing in many contexts. Likewise, although most clients are men, consumption of sex is certainly not limited to male clients, as evidenced by the increasing presence of female sex tourists in developing countries, especially in the Caribbean region (Sanchez Taylor 2001; O'Connell Davidson & Sanchez Taylor 2005).

Places where sex is sold can include streets and highways, brothels, hotels, bars, night clubs and strip clubs, massage parlors, restaurants, discotheques, saunas, private apartments, dungeons for bondage and domination services, websites, call girl and escort agencies, phone sex businesses, and private parties for erotic and fetish services. The organization of the industry will vary from context to context, depending on the social and legal structures that are in place.

Many businesses test the limits of legality, as when a bar licensed to sell alcohol also sells sex occasionally, or when legally registered dungeons organize informal parties where sex is sold. As it has diversified, the sex sector has also become more and more enmeshed in other economic sectors. This symbiotic relationship is present in, for example, profits to real estate vendors from rent paid by sex-related businesses and the sale of pornographic videos by hotel or media corporations.

Expanding demand for sexual services is fueled both by demographic shifts and new technologies. A declining marriage rate in many developed countries, and an increase in the number of persons living alone add to demand for erotic services (Bernstein 2005). Technological innovations have also fueled the rapid expansion and diversification of the commercial sex sector around the world over the past 15-20 years, consolidating what many estimate now to be a global, multi-billion dollar industry. The introduction of home video technology in the early 1980s is credited with much of the boom in the pornography industry, and video and Internet pornography have now almost entirely replaced adult theaters. The Internet and cellular telephones have allowed for businesses to become more mobile and private and, in places where the sale of sex is criminalized, often to escape detection by law enforcement. For example, private entrepreneurs who organize sex or fetish parties may change location from week to week, announcing the new venue via email to a pre-screened list of participants and requiring a password for entry. Similarly, mobile phones allow more commercial sex workers to operate independently, as when call girls develop their own list of clients and work on a freelance basis.

The expansion of the commercial sex industry, especially in developed countries,

has been theorized by some as a means of compensating males for the loss of power they have experienced both as a result of the gains of second-wave feminism (O'Connell Davidson 1998) and women's incursion into the labor market and economic realms (Kimmel 2000). More recently, Bernstein (2005) has challenged the assumption, implicit in these theories, that sexual desires might preferably be fulfilled within an established relationship, and posited additionally that commercialized sexual services may provide elements of intimacy that the household cannot, or at least is not, offering. These new erotic possibilities need not threaten the stability of the nuclear family, and may even be complementary: ethnographers have revealed how happily married men seek out extramarital sexual services to satisfy a curiosity or need for variety, rather than to compensate for something lacking at home (Monto 2000, Prasad 1999).

Consumers increasingly look to the marketplace to provide these services, many of which are specifically intended *not* to feel like a stereotypical prostitute-client encounter. Although money is exchanged, these transactions take on many characteristics of normalized intimate, even romantic, relationships. The so-called "Girlfriend Experience" marketed by advertisements across the United States is one example of these services. The GFE, as it is abbreviated, is intended to provide the feeling of a relationship within the temporal and physical boundaries of an economic transaction.

In this way, the separation of intimacy from romantic love is normalized and facilitated by its location in the commercial market (Bernstein 2005). In the process, commercial sex takes on the characteristics of other commodities. Among these: sexual services are available not only to an elite class but to

society more broadly; prohibition (criminalization) has minimal effect on the quantity exchanged; the quality of the service determines whether the exchange will continue; and social norms influence the type of services offered (Prasad 1999:188).

Economic and Social Organization

The underground nature of much of the sex industry often obscures its integration with other economic sectors and broader social life. In reality, the illegal or marginally legal aspects of the industry are often tangled up with established, formal businesses and rely on the support of formal businesspeople, politicians, police, lawyers and others. Despite these links, however, the sex sector remains isolated from other aspects of political and social life, both in research and policy. Researchers rarely include sex workers in studies of the informal economy, migration, or labor rights, for example. Instead, studies focusing on the sale of sex tend to examine only the transaction between client and sex worker, rather than the context in which the transaction takes place (Agustín 2002). Policymakers usually address the sex sector through moral and legal systems, either by criminalizing or attaching a social stigma to it (Lim 1998). As discussed below, this artificial exclusion of the sex sector from other areas of social and economic activity has often contributed to ineffective and contradictory policies.

The foundations of the sex industry are undoubtedly economic, and economic forces also drive the expansion and entrenchment of the sex sector today. In some developing and middle-income countries, this is a product of the increasing mobility and materialism that have accompanied economic development. In Malaysia, for example, the International Labour Office has documented how rising incomes and purchasing power have fed a largely male demand for sexual services (Lim

1998). These services often form part of the business culture, as when business meetings are conducted at hospitality clubs, entertainment clubs or other commercial sex venues. This is consistent with research on client demand in rich countries, which has shown that [male] clients purchase more commercial sex as their age and income increase, and when they hold steady employment (Giusta et al 2006). On the supply side, although the increased educational opportunities for women that accompany economic development might be expected to decrease the number of women willing to enter the sex sector, these opportunities have not always been matched by alternative employment prospects sufficient to deter women from entering (Lim 1998). Thus, the rising income levels that result from economic development have often led to the growth of the sex sector rather than its decline.

In poorer countries, growth of the sex sector often forms part of national development policy, although it is rarely explicitly mentioned. The sex sector can play a direct role in development strategies when countries place a growing emphasis on tourism. The growth and diversification of the sex sector in developing countries to attract foreign clientele has been documented in various contexts, especially in the Caribbean region (O'Connell Davidson and Sanchez Taylor 2005; Kempadoo 1998, 2004; Padilla 2007; see also Lim 1998 on Southeast Asia and Troung 1990 on Vietnam). In other cases, the role of the sex sector is more implicit, as a way to mitigate problems of unemployment, declining incomes and rising costs of living. Structural reforms, often mandated as part of external debt repayment plans, have focused on reducing state employment, devaluing the currency and trimming social services. These have often had the combined effect of increasing the cost of living by reducing

access to services and putting downward pressure on wages. These trends place a disproportionate burden on women to cushion the effects of economic crisis and reform, as they are increasingly expected to contribute to family income even as their household responsibilities remain constant. For many women, especially single mothers, sex work combines with other, mostly informal income-earning activities to enable them to balance economic and family responsibilities. Indeed, single mothers tend to account for a large proportion of the sex worker population in various contexts (Sanders 2005a; Lim 1998; Brennan 2004; Willman-Navarro 2008). Sex workers can often obtain higher earnings in less time than other low-skill workers in the formal manufacturing and service sectors.

The social roots of prostitution have always been based in societal institutions that define relationships between men and women, and between parents and children. In patriarchal systems, men are allowed freedom to define their own rules in terms of sexuality, including access to various forms of sexual pleasure and with multiple [female] partners (D'Cunha 1992: 38). Women, in contrast, are expected to confine their sexuality to heterosexual, monogamous relationships, or risk being categorized as deviant. The transgression of prescribed social norms for the "good" (virgin, sexually pure) woman, leads to condemnation and marginalization as "bad" (whore, promiscuous). Historically, prostitution has been considered a 'necessary evil,' whereby prostitutes serve as an outlet for uncontrollable male desire that must exist to protect innocent women, thereby maintaining the stability of the heteronormative nuclear family. This premium on virginity and sexual purity often influences women's entry into prostitution after a rape, abandonment by a male partner, or other extra-marital sexual experience that

labels them as "ruined" for marriage (Lim 1998).

This virgin/whore dichotomy also generates the most powerful factor behind the social organization of the sex industry - the social stigma attached to commercial sex. This stigma keeps sex workers and the sex industry artificially isolated from broader economic and social life. Because they are labeled as 'deviant,' sex workers who are identified as such may have difficulty securing other work, finding romantic partners, and interacting with broader society in general. In addition to these emotional stresses, stigma can contribute to more extreme vulnerability, as when sex workers are reluctant to seek police support after a violent encounter with a client, or to access needed health services. To mitigate this stigma, sex workers often employ sophisticated strategies to avoid being identified as a sex worker, including migrating to another city or country for work (Lim 1998; Brennan 2004; Belsey 1996) or lying to family or partners about their sources of income (Willman-Navarro 2006). Sanders (2005b) has argued that the emotional stresses associated with the fear of being "found out" are among the top concerns sex workers face. The social stigma attached to sex work is not limited to sex workers; clients also often worry about being "caught" purchasing sex, and may seek environments that are more private or exclusive in order to manage this.

Family relationships also play a large role in influencing the social organization of the sex sector. In some Asian countries, the cultural notion of filial piety often implies a moral obligation to compensate parents for the care received in childrearing. Under increasing economic pressures, the role of daughters within the family has been changing, and more and more young women are expected to contribute to family income. In cases where women migrate to work in the

sex industry and send remittances, these pressures conflict with traditional norms of the pure, chaste woman. If they are able to bring economic security to their families, migrant sex workers often return home with honor, despite having obtained the earnings through stigmatized activities (Belsey 1996). Other cultural norms have contributed to the most exploitative parts of the sex sector, as when family expectations and economic pressures push children into prostitution. The sexual exploitation of children is also driven by cultural myths that claim sex with virgins will restore virility or even cure sexually transmitted diseases.

The influence of religion on the organization of the sex sector is somewhat ambiguous. Although all major religions condemn commercial sex, no evidence substantiates the theory that prostitution is less prevalent in countries governed by religious law. In some countries, such as Indonesia and Malaysia, prostitution is illegal under religious (*Shariah*) law, but not criminal law. Thus, religion does not appear to have a strong relationship in either increasing or reducing supply or demand of commercial sex (Lim 1998).

Size and Significance

The size and significance of the sex sector, as well as its relationship with other economic sectors, are difficult if not impossible to measure given the underground nature of much of the industry. Few governments collect and publish data on the sex sector even where most activities are legalized and regulated. Some estimates on various countries by non-governmental organizations using mapping and survey methods are given in table 1. Government statistics on the UK estimate that 80,000 sex workers are employed in the sex industry (Home Office 2004). In Germany, the sex industry is estimated to be worth 14.5 billion euros

annually and employs approximately 400,000 workers (Mitrovic 2004). In the US, the pornography industry alone brings in an estimated \$10 billion a year - more than the NFL, NBA and the Major League combined - and has attracted investment by companies from General Motors to Marriott and Time Warner (Schlosser 2003).

In developing countries, statistics are unreliable or nonexistent, but some basic estimates exist. The ILO survey covering Indonesia, the Philippines, Malaysia and Thailand estimated that the sex sector accounts for between two and 14 percent of Gross Domestic Product, and employs between 0.25 percent and 1.5 percent of the female population in these countries (Lim 1998).

Despite the lack of systematic data collection, there is some consensus in government, scholarly and media circles that on a global level the sex industry is growing and diversifying, and in the process, further integrating itself into the overall economy. These affirmations are common in the literature on human trafficking (Bales 2005), globalization (Altman 2001; Sassen 2002), feminist theory (Barry 1995) and migration (Agustin 2005). Indeed, the same forces of globalization that have strengthened formal investment and communication networks have also reinforced a web of informal and illicit networks through which the sex industry often operates.

Labor migration has been an important force in expanding the size of the global sex industry. In Europe, independent organizations estimate that 70 percent of sex workers in the European Union are migrants, many of them without formal legal status. Mapping exercises to measure the population have found that over the last decade, the number and diversity of sex workers have increased. In 1995, ten to twelve nationalities were represented among European sex

workers, compared to 40-45 different nationalities in 2002-04. Women from Central and Eastern European countries are estimated to account for 30-40 percent of these (TAMPEP 2004). Some migrant sex workers work on a seasonal basis by entering with tourist visas for periods of 3-6 months at a time. In more extreme cases they may be coerced by traffickers and effectively prohibited from returning to their countries of origin through debt bondage. (See further discussion in working conditions, below.)

Commercial Sex and the Service Economy

Two competing discourses dominate the research in social policy, feminist and governance research: that of prostitution (and pornography) as violence against women, and that of sex work as work (O'Neill 2001). Much of the controversy centers on the question of whether the activities of women who sell sex constitute a type of sexual and emotional labor within the broader service economy, or those activities represent women's victimization and objectification within a patriarchal, capitalist system that subordinates female bodies to male domination (for a review see Chapkis 1997).

At one end of the spectrum, what can be termed the anti-prostitution, or abolitionist feminists, position commercial sex within an oppressive patriarchal structure that holds men and women to different standards of morality. From this view, women cannot consent to selling sex or participating in pornography - these activities always constitute violence (Barry 1995, Jeffreys 1997, Pateman 1998). A woman cannot be a *sex worker*, because within the context of unequal power relations between men and women, the act of selling sex automatically converts her into a *sex object* (Mackinnon 1987; for an overview see Sanders 2005b). According to this theory, the harm inherent in selling sex is that in selling the client an

objectified female body, the prostitute or pornography actress sells *herself* in a deeper sense than that associated with other jobs (Pateman 1988). This literature focuses on prostitution and pornography as forms of violence and advocates for policies that help women leave the sex industry and punish clients (see section on governance for more on these end-demand policies).

The "sex work as work" movement emerged as part of an effort to define the sale of sex as a legitimate form of labor deserving of the same protections that characterize other professions, and to highlight the connections between sex workers and other workers, especially in the low-skill service sector (Leigh 1997; Kempadoo and Doezema 1998). From this perspective, sex workers can sell sexual services without selling *themselves*, and thus has much in common with other service professions. Chapkis (1997), for example, has connected the notion of sexual labor to emotional labor, a sociological category describing forms of labor that embody caring and sentiments associated with intimate relationships in professions such as waiting tables, nursing, childcare, flight attendants and other service jobs. These [feminized] professions carry with them the performance of femininity an expectation for employment, and female sexualization as part of the work culture (see Sanders 2005b for a review). Like other service workers, they argue, sex work involves the sale of a service, not the self, and sex workers often employ complicated emotional and identity management strategies to maintain boundaries between themselves and clients. Common strategies include using the condom as a psychological barrier, or limiting client contact to certain regions of the body (Sanders 2002, 2005a; Chapkis 1997).

Within the "sex work as work" perspective, some view the sale of sex not only as a legitimate form of labor, but also a challenge

to oppressive social norms (see O'Connell Davidson 2002 for a discussion). They locate the roots of oppression in the way that "society assigns privilege based on adherence to moral codes" (Califa 1994: 11) rather than patriarchy in general. By selling sex, women subvert the very patriarchal system that oppresses them (Pheterson 1993, McClintock 1993). As Kempadoo and Doezema (1998) have argued, the structure of the sex industry is highly gendered, positioning "the socially gendered category 'women' as the sellers or providers of sexual labor and 'men' as the group deriving profits or power from the interactions" (1998:5). This arrangement often prevails even in the case of male sex workers servicing male clients, as the male sex worker often performs a feminized role in the transaction. This does not imply, however, acceptance of the sale of sex simply as a confirmation of male domination. According to Chapkis (1997), sex work involves "sites of ingenious resistance and cultural subversion...The prostitute cannot be reduced to one of a passive object used in male sexual practice, but instead it can be understood as a place of agency where the sex worker makes use of the existing social order" (1997: 29-30). Indeed, they often point out that the sex industry is one of the few economic areas where women can earn more than their male counterparts. Prostitution is power—a woman's ability to use her sexuality to manipulate the power structures that place her at an economic disadvantage. It is an act of protest that sex workers are "demanding, and generally getting, better money for [their] services than the average, male, white-collar worker" (McClintock 1993: 33).

In addition to the overarching gendered power relations, race and class relations also define the organization of the sex industry. The eroticization of certain races and cultures simultaneously places a sexual value and a social inferiority on particular races,

especially in the Third World. Trends such as sex tourism are centered on notions of women and men in developing countries as sexual objects, available to mostly white male sex tourists. On a global level the sex industry is structured around a racial hierarchy that positions white women at the top, able to work in mostly protected, indoor environments; Mulatto, Asian and Latina women as a middle class; and Black women in the most vulnerable conditions, especially street work (Kempadoo and Doezema 1998; Chapkis 1997). By highlighting these hierarchies, some within the sex work as labor movement aim to link the struggles of sex workers to those of workers in other exploitative sectors, such as informal and unregulated work, creating the potential for international solidarity among women and workers in general (Mohanty 1997; Kempadoo and Doezema 1998). Indeed, these analysts have asserted that the practice of sex work is sometimes understood as a form of resistance to these unequal gender, race and class relations, an argument that those advocating for the rights of workers in other sectors rarely use (O'Connell Davidson 2002).

Others, while sympathizing with the need to understand and protect sex work as a form of labor, are reluctant to celebrate it as resistance. They have highlighted the ways that consent and subversion are shaped by racial, gender and class inequalities. Julia O'Connell Davidson and Sanchez Taylor (2005) argue that "individuals do not give consent (whether to a system of political governance, a wage-labor contract, or to sexual interaction) in a social, political and economic vacuum, and consent cannot be meaningfully abstracted from the power relations that surround it." Research on sex tourism in the Dominican Republic, for example, has shown that although sex workers may be viewed as exploiting the

foreigners who come to exploit them, in reality few manage to capture the coveted visa or economic stability they originally envision (Brennan 2004).

It is within this growing middle ground that a number of empirical studies have emerged in recent years, illuminating the diversity of experiences in the sex sector. This has included research into male clients of female workers (Monto 2000; Giusta et al 2006) and of male sex workers (Padilla 2007); female clients of male sex workers (Sanchez Taylor 2001; O'Connell Davidson and Sanchez Taylor 2005); and the economic dimensions of risk and earnings (Gertler et al 2005; Willman-Navarro 2008), as well as the sociological construction of risk (Sanders 2005a). They have also expanded discussion of sex work beyond street prostitution to indoor settings (Sanders 2005a, Willman-Navarro 2008) and forms of sex work beyond "prostitution", especially stripping (Wood 2001; Montmurro et al 2003).

Despite these explorations, there remains a paucity of theorizing on several dimensions of the sex sector, especially regarding financial psychology of sex workers (Weldon 2006), interpersonal relationships, migration, culture and others. Laura Agustín advocates a broader, 'cultural approach' to the study of sex work that can include activities of both commerce and sex. That is, addressing the intersections of commercial sex with broader cultural concerns including, "art, ethics, consumption, family life, entertainment, sport, economics, urban space, sexuality, tourism and criminality, not omitting issues of race, class gender, identity and citizenship" (2005: 682).

Working Conditions and Exploitation

Generally speaking, the sale of sex within the sex industry falls into two broad structural categories: organized and unorganized. Organized arrangements rely on an owner or

manager and involve defined relationships between management and sex workers. They may often involve the participation of an intermediary, who puts clients in touch with providers of erotic services. These can include pimps and madams, owners and managers of sex industry establishments such as clubs, brothels, bars, hotels or massage parlors or the staff of escort agencies. Unorganized arrangements are those where sex workers seek and solicit clients independently, although they may rely on other actors for protection (Lim 1998). Working conditions in different types of arrangements vary enormously and can range from slave-like and abusive to comfortable and lucrative.

Much of the media and popular literature tends to focus on the more exploitative and repugnant aspects of the sex industry, especially forced prostitution and trafficking. Definition and measurement of the trafficking problem has been immensely challenging. Arriving at an internationally agreed definition of human trafficking has been difficult and was only obtained in 2000 through the signing of the United Nations Protocol to Prevent, Suppress and Punish Trafficking in Persons in December 2000 (United Nations 2000). Articles 2-3 of the Protocol define human trafficking as the "... recruitment, transportation, transfer, harbouring or receipt of persons..." by force, abduction, fraud or coercion and for improper purposes such as forced labor, slavery, or sexual exploitation. Signatory countries must commit to passing national legislation penalizing these activities and protecting individuals who have been trafficked.

Estimates of the number of people trafficked are notoriously problematic and vary enormously depending on the source. Many governments and organizations often do not distinguish between individuals trafficked for sexual exploitation from those

trafficked for other forms of forced labor, such as agricultural or sweatshop labor. For example the US government's annual report on trafficking in persons estimates that between 600,000 to 800,000 people are trafficked each year, but includes those trafficked for all types of forced labor (US Department of State 2005). Other international organizations are more reluctant to report statistics, given past criticism of the unreliability of the numbers. To tackle these measurement issues, UNESCO has established a Trafficking Statistics Project that evaluates the validity of statistics from different official sources (UNESCO 2006).

From the “prostitution as violence” perspective, coercion and trafficking are inherent to the sex industry. The anti-trafficking, anti-prostitution organization Coalition Against Trafficking in Women, a leading advocacy group, rejects all forms of commercial sex as exploitative, regardless of consent, and inextricably linked to trafficking. (CATW 2006 and Raymond 2005). These groups claim that legalized or tolerated commercial sex markets facilitate trafficking and child prostitution.

Sex work as labor advocates counter that exploitation and coercion are not specific to the sex industry, but rather are a product of exploitative working conditions caused by the artificial isolation of the sex sector from other economic sectors. The marginalization of sex workers, they argue, inhibits their ability to report abuse and seek protection under the law. Perhaps the most important aspect of fair and acceptable working conditions is the ability to discriminate among clients; that is, the freedom to decline both clients and/or specific services at the sex worker's discretion (Bindman with Doezema 1997). The right to determine one's working conditions is also a central component of the World Charter of Prostitutes' Rights, adopted by sex worker rights organizations in Amsterdam in 1985.

The Charter stresses that it is “essential that prostitutes can provide their services under the conditions that are absolutely determined by themselves and no one else” (ICPR 1989: 40) Such conditions include the freedom to choose one's place of work and residence (unimpeded by systematic zoning laws or selective geographic policing of prostitutes), and freedom from restrictions on travel within and between countries.

Some of the most successful interventions to improve working conditions have been undertaken with the participation or leadership of sex workers themselves. When sex workers are legally empowered to report abuse they can be strong allies in combating trafficking. The Sonagatchi Sex Worker Intervention Project in India is often touted as one successful example of employing sex workers as peer educators to identify and report situations of abuse and coercion, assist in helping victims of trafficking and child prostitution, and implement health and safety education campaigns.

Governance Issues

Governance of the sex sector involves consideration of a broad range of concerns including human rights, employment, income distribution, gender inequality, and the exploitation of women and children (Lim 1998). Many of the dilemmas of governing the sex sector stem precisely from the range of experiences of those who are involved in it. The method of entry is especially important in defining these experiences - some are directly coerced and trafficked while others are pushed by economic circumstances or family obligations to enter. Still others choose sex work as a means of obtaining higher earnings or migration opportunities. The resulting employment conditions can therefore range widely from exploitative and violent, to comfortable and well remunerated. Recognizing this

complexity, many international organizations, including the International Labour Office, do not take an official stand on the governance of the sex industry (Lim 1998); however, international organizations and most national governments do share some legal stances regarding the most exploitative aspects of the industry. For example, pimping, trafficking and child prostitution are illegal in most countries, although interpretation and enforcement of these laws varies widely.

Policies addressing the sex sector focus primarily on individuals who sell sexual services, rather than on regulating establishments. That is, governance is almost entirely centered on regulating (or prohibiting) the commercial sex transaction between sex worker and client, leaving the surrounding business establishments largely unaffected (Agustin 2005). In many cases, businesses within the sex sector operate informally, so that their activities are not recorded in official accounting and health, safety, tax and worker protection legislation is not enforced. In other cases, businesses where sex is sold may be only partially regulated, by holding legal license to sell alcohol or food, but not to sell sex. In still other contexts, bribes to public officials and police allow establishments to avoid government regulation and inspection. A smaller number operates completely off the books, avoiding detection by law enforcement and, when detected, often changing locations to continue doing business.

Commercial sex establishments often take advantage of legal gray areas for profit. For example, many strip clubs in the United States offer lap dancing, which occupies an ambiguous space between performance, which is legal, and prostitution, which is not (Chapkis 2000). Businesses may also circumvent anti-discrimination and other labor laws by employing workers on a contract basis and requiring workers to call in

to request a shift each day, allowing them to allocate shifts based on racial preferences, exclude workers trying to organize, and other unfair labor practices (Kempadoo 1998).

With regard to prostitution specifically, there are three dominant approaches to governance common at the national level: criminalization, decriminalization and legalization. *Criminalization* of prostitution involves the penalization of different aspects of the sex industry. This may take the form of complete *prohibition* of all activities related to commercial sex, including soliciting to sell or buy sex, pimping, owning a brothel or otherwise living off of, or benefiting from, the earnings of prostitution. This is the model for much of the United States, Thailand, and some Gulf states. In other countries, the sale and purchase of sex is not criminalized, but related activities are, including advertising, recruiting and operating commercial sex establishments. This is the framework for most of Western Europe, the United Kingdom, Canada, parts of Australia, some countries in Southeast Asia and most of Latin America.

Sex workers' rights organizations object to any form of criminalization on the grounds that it increases the vulnerability sex workers face (Declaration of the Rights of Sex Workers in Europe 2005). If they fear arrest, sex workers may be more reluctant to report abuses by clients and employers, and may be forced to work in unsafe, underground environments or enlist the protection of pimps and other intermediaries. On the street, the fear of arrest limits the time sex workers can spend negotiating with clients and identifying security concerns, such as assessing whether a client is drunk or drugged (Barnard 1993; Sanders 2005a). In addition, the prospect of arrest makes sex workers more vulnerable to police harassment and abuse, one of the most significant concerns of sex workers in many countries (*Research for Sex Work*, No. 8,

2005). A further concern is that even in contexts where prostitution is legal, street workers may be arrested and prosecuted if they are migrants. In France, the 2003 Sarkozy law was adopted against the opposition of organized sex workers, which allows police to arrest and deport migrant street workers, driving many of them further underground to more vulnerable environments.

Sex worker rights organizations also criticize laws that criminalize profiting from commercial sex by intermediaries because such laws restrict sex workers' autonomy in ways not experienced by workers in other sectors. Where brothels are illegal, laws may prohibit two sex workers from working out of the same private apartment. In other cases, laws against pimping can be applied to the live-in partners or even adult children of sex workers, who may be required to prove financial independence, or even landlords and others who receive part of her earnings (Mikulski 1993; Declaration of the Rights of Sex Workers in Europe 2005).

Criminalization strategies have traditionally focused on the arrest of individual prostitutes, particularly those working on the street. However, rather than reduce the overall incidence of prostitution, criminalization usually works to contain activity within a particular area, or displacing it to other, often indoor sites with less police interference. This approach reflects a general opposition to prostitution as an environmental nuisance, rather than on moral grounds per se (see Weitzer 2000 for analysis on the United States).

Over the last 15 years, governance in some Western countries, notably the US and the UK, has shifted from policing to a combined law enforcement and welfarist approach, whereby various social service agencies assist sex workers in leaving prostitution. Some have argued that by coupling criminalization

with rehabilitation programs, the law victimizes sex workers further, as it "locates individual women as being responsible for the social problems they encounter, thereby justifying a punitive response, when, despite the best efforts of support agencies around them, they continue with their involvement in prostitution" (Phoenix and Oerton 2005:100). These approaches have been indirectly extended to developing countries through restrictions on official US aid to organizations that provide any services to sex workers other than those aimed at helping them leave prostitution. Two lawsuits were filed, by the Open Society Institute in 2005 and by HIV service provider DKT International in 2006, against the US Agency for International Development recently challenged these restrictions.

Abolitionist groups also object to criminalizing the sale of sex, arguing that in practice, it often ignores the role of the client in the transaction. Indeed, feminists of all stripes have long criticized the disproportionate treatment of prostitution in the criminal justice system that arrests women and sends male clients home without consequence (Pheterson 1993; Jeffreys 1997). This has led to the emergence of "end-demand" initiatives, many of which are organized on the community level and are aimed at shaming clients away from purchasing sex. In the United States, these have included placing photos of alleged or convicted clients on billboards or the Internet, at times in coordination with local police departments. Some police departments have also begun confiscating the cars of men picked up for soliciting. In Chicago, offenders must pay a fine of between \$750 - \$1,500 to retrieve their cars.

Perhaps the fastest growing end-demand initiatives are "John Schools", diversion programs for men arrested for soliciting commercial sex. The projects usually consist

of a one-day seminar offered as an alternative to public trial, aimed at educating men about the harmful effects of prostitution on individuals and communities. There are now at least 12 John school prostitution diversion programs operating in the United States and 14 in Canada. (For analysis of John School programs in Canada and the United States, see Monto and Garcia 2001; Wortley Fischer 2002; Bernstein 2005).

Sweden has taken a novel approach to ending demand by criminalizing the purchase of sex, while decriminalizing prostitution. Under the "Swedish model", sex workers (assumed to be primarily women) are free to offer sex, but [male] clients are prohibited from buying it. The arrest of clients is meant to correct the double standard in law enforcement and by reducing demand, decrease the number of women involved in prostitution. Of three government reports commissioned to evaluate the program's success, however, none concluded that the law has resulted in a significant drop in prostitution. To the contrary, the number of bordellos appeared to have increased. Social workers and sex workers claim that the quality of the clientele has declined, as those worried about arrest have gone indoors, leaving only the riskiest on the streets. They also claim police harassment has increased because women can be forced to appear in court to testify against clients. An especially alarming, unintended consequence of this approach is that because condoms can be used as evidence in the clients' trials, clients have a strong disincentive to use them (see Kulick 2003 for a full analysis of the law and its complications).

Decriminalization is meant to remove all criminal penalties on commercial sex and leave prostitution unregulated. Taken to an extreme, decriminalization would mean prostitution could be practiced in any area without restriction, and workers and

businesses would not be required to comply with the regulations required of other commercial industries (Weitzer 2000). Sex workers' rights organizations have strongly advocated for decriminalization for several reasons. First, decriminalization could encourage a shift in law enforcement efforts from policing sex workers to protecting them. They also claim that the repeal of anti-pimping laws would allow sex workers to organize businesses together, and correct the law enforcement practice of using these laws against their partners and managers. Critics counter that removing regulations gives the sex industry advantages not enjoyed by other economic sectors.

No country has enacted full decriminalization of the sex sector. In 2003, New Zealand enacted a reform of prostitution policy that decriminalizes soliciting, makes it legal to live off of earnings from prostitution, and allows sex workers to take clients to court if they do not pay. The system is not without regulation, however: all indoor establishments were brought under a brothel-licensing regime subject to tax regulation and health and safety legislation, including a requirement to ensure safe sex practices. In addition, local authorities have the power to decide where brothels are located and ban advertising for prostitution in public places.

Legalization can describe a broad range of legal frameworks, including contexts where prostitution is simply not addressed in the law and is therefore legal by default. More often, it implies some form of government control and regulation, which can take a number of forms. This may include licensing, required health exams, government owned or licensed brothels, special tax regulations for sex workers, and zoning laws for the location of sex-related businesses, sex worker residences and outdoor prostitution. Legalization is usually adopted to better monitor prostitution,

decrease exploitation of individual prostitutes and protect public health.

Critics of legalization, including many sex worker organizations, claim that legalization causes more problems than it solves. Anti-prostitution groups argue that legalized prostitution simply sanctions the exploitation and abuse inherent in the industry. Other critics argue that regulations are often unfair and even counter-productive. These maintain that health checks are an intrusive and impractical attempt to control women's bodies, and that requiring registration by sex workers separates those who are comfortable disclosing their identity as a sex worker from those who are not. In addition, by separating sex workers from other professions in this way, they claim legalization systems only reinforce the stigma sex workers face, limiting their opportunities for exit and transition to other jobs (Soothill and Sanders 2004). Another issue is that even where prostitution is legal, it is difficult to ensure that sex workers will comply with the laws, especially regarding street prostitution. In Nevada, for example, street prostitution is illegal but rampant in Las Vegas, despite the existence of government-owned brothels in nearby counties. Another concern with legalization is that it might encourage more people to enter prostitution; however, this has not been the case in most places where prostitution is legalized, including The Netherlands (Weitzer 2000:177).

In 2002, Germany adopted a form of legalization still unique in the world, by placing regulation of sex work under labor law. That is, sex workers are subject to the same regulations and protections other workers enjoy, including tax collection and access to public health and social security services, and brothels are regulated just as other businesses are. Street prostitution - generally illegal everywhere - is legal in Germany, but cities and provinces often

restrict it to certain areas through zoning regulations.

Regardless of the formal framework for governance of the sex industry, however, state *tolerance* tends to be a key component of policy. Tolerance can be expressed in the selective enforcement of the law - for example by limiting policing to outdoor prostitution or specific geographic areas - or by the absence of a general law regarding commercial sex. In countries where commercial sex is not explicitly legal or illegal, it may be indirectly regulated and contained using other aspects of criminal law, including public disorder or vagrancy laws (Weitzer 2000).

Sex workers, arguably much more than workers in other industries, have traditionally been excluded from public debates that affect their working conditions. In Sweden, since the 2003 law criminalizing the purchase of sex, it has become common for politicians to threaten to withdraw from public debates in which sex workers were invited to participate (Declaration of the Rights of Sex Workers in Europe 2005). For this reason sex worker rights organizations, including unions where they exist, have advocated for a stronger voice in policy debates. There has been little success with this on the national level generally, but on the neighborhood level, some initiatives have emerged that involve sex workers in the research and planning of new policies. These efforts aim to bring the concerns of community members and of sex workers together to ensure more effective service delivery to vulnerable populations, and to better address the environmental concerns of communities. In a series of such projects in Midlands, UK, researchers identified needs for improved services in housing, drug rehabilitation, and mediation services between community members and sex workers that resulted in increased collaboration among agencies at a local level

(O'Neill and Pitcher 2006). Such initiatives hold promise for resolving conflicts related to commercial sex and to "establish ways to introduce effective urban policy that is neither punitive, moralistic nor biased" (Sanders 2004:1715).

Conclusion

The global sex sector has rapidly expanded and diversified with globalization, further integrating with other economic sectors and with social institutions. This has given rise to new issues in governance centering mostly on a debate over whether commercial sex is inherently exploitative, or a form of labor to be protected and regulated. Social norms often contribute to the stigmatization and marginalization of those involved in the sex industry. In most contexts, policy remains primarily morally based, with policymakers continually reluctant to address the social and economic bases of the sex sector. The challenge for policy is to consider the sex sector through the lenses of broader concerns including employment, income distribution, gender discrimination and human rights in an effort to address more effectively its exploitative elements.

Selected References

- Agustín, Laura Maria. (2005) "The Cultural Study of Commercial Sex", *Sexualities*, 8, 5, 618-631.
- Altman, Dennis. (2001) *Global Sex*. Chicago: IL: University of Chicago Press.
- Bales, Kevin. (2005) *Understanding Global Slavery: A Reader*. Berkley and Los Angeles: University of California Press.
- Barnard, Marina (1993) "Violence and Vulnerability: Conditions of Work for Street Working Prostitutes". *Sociology of Health and Illness*, 15, 1, 5-14.
- Barry, Kathleen. (1995) *The Prostitution of Sexuality*. New York: New York University Press.
- Belsey, John. (1996) *Commercial Sexual Exploitation of Children: Health and Psychosocial Dimensions*. WHO paper presented to the World Congress Against the Commercial Sexual Exploitation of Children, Stockholm, Sweden, 27-31 August.
- Bernstein, Elizabeth. (2005) "Desire, Demand and the Commerce of Sex", in Elizabeth Bernstein and Laurie Schaffner (Editors), *Regulating Sex: The Politics of Intimacy and Identity*. New York & London: Routledge.
- Bindman, Jo and Jo Doezema. (1997) "Redefining Prostitution as Sex Work on the International Agenda". www.walnet.org/csis/papers/redefining.html
- Brennan, Denise. (2004) *What's Love Got To Do With It? Transnational Desires and Sex Tourism in the Dominican Republic*. Durham and London: Duke University Press.
- Campbell, C. (1991) "Prostitution, AIDS and Preventative Health Behaviour", *Social Science and Medicine*, 32, 12, 1367-1378
- Campbell, C. (1997) "Migrancy, Masculine Identities and AIDS: The Psychosocial Context of HIV Transmission on the South African Gold Mines", *Social Science and Medicine*, 45, 2, 273-81.
- Chapkis, Wendy. (2000) "Power and Control in the Commercial Sex Trade", in Ron Weitzer (Editor), *Sex for Sale: Prostitution, Pornography and the Sex Industry*. New York and London: Routledge.
- Declaration of the Rights of Sex Workers in Europe*. (2005) Document adopted by the European Conference on Sex Work, Human Rights, Labour and Migration. Brussels, October.
- Gertler, P., M. Shah and S. Bertozzi. (2005) "Risky Business: The Market for Unprotected Sex", *Journal of Political Economy*, 113, 3.
- Giusta, Marina Della; Maria Laura Di

- Tommaso; Isilda Shima and Steinar Strøm. (2006) *What Money Buys: Clients of Street Sex Workers in the US*, Memorandum No. 10, Department of Economics, University of Oslo.
- Home Office. (2006) *A Coordinated Prostitution Strategy and a Summary of Responses to Paying the Price*. London: UK Home Office, Ref 272138
- Jeffreys, S. (1997) *The Idea of Prostitution*. North Melbourne, Australia: Spinifex.
- Kaiser Network. (2005) *Daily HIV/AIDS Report*. www.kaisernetwork.org/
- Kaye, Kerwin. (2007) "Sex and the Unspoken in Male Street Prostitution", *Journal of Homosexuality*, 53, 1, 37-73.
- Kempadoo, Kamala and Jo Doezema. (1998) (Editors) *Global Sex Workers: Rights, Resistance and Redefinition*. New York and London: Routledge.
- Kempadoo, Kamala. (1998) "The Exotic Dancers Alliance: An Interview with Dawn Passar and Johanan Breyer", in Kamala Kempadoo and Jo Doezema (Editors), *Global Sex Workers: Rights, Resistance and Redefinition*. New York & London: Routledge.
- Kempadoo, Kamala. (2004) *Sexing the Caribbean: Gender, Race and Sexual Labor*. New York: Routledge.
- Kimmel, M. (2000) "Fuel for Fantasy: The Ideological Construction of Male Lust", in Kerwin Kaye, Baruch Gould, and Jim Nagle (Editors), *Male Lust: Power, Pleasure and Transformation*. New York: Hawthorne, pp.267-273.
- Kulick, Don. (1998) *Travesti: Sex, Gender and Culture Among Brazilian Transgendered Prostitutes*. Chicago: University of Chicago Press.
- Kulick, Don. (2003) "Sex in the New Europe: The Criminalization of Clients and Swedish Fear of Penetration", *Anthropological Theory*, 3, pp.199-218.
- Leigh, Carol. (1997) "Inventing Sex Work", in Jill Nagle (Editor), *Whores and Other Feminists*. London: Routledge, 223-231.
- Lim, Lin Lean. (1998) (Editor) *The Sex Sector: The Economic and Social Bases of Prostitution in Southeast Asia*. Geneva: International Labour Organisation.
- MacKinnon, C. (1987) *Feminism Unmodified: Discourses on Life and Law*. Cambridge, MA: Harvard University Press.
- McClintock, Anne. (1993) "Sex Workers and Sex Work: Introduction", *Social Text*, 37, Winter, 1-10.
- Mikulski, Barbara A. (1993) "It's a Pleasure Doing Business With You", *Social Text*, 37, Winter, 11-22.
- Mitrovic, Emilija. (2004) *Working in the Sex Industry: Report on the Findings of a Field Research*. Ver.di. Berlin.
- Mohanty, Chandra Talpade. (1997) "Women Workers and Capitalist Scripts: Ideologies of Domination, Common Interest and the Politics of Solidarity", in M. Jaqui Alexander and Chandra Talpade Mohanty (Editors), *Feminist Genealogies, Colonial Legacies, Democratic Futures*. New York: Routledge.
- Montemurro, Beth; Coleen Bloom and Kelly Madell. (2003) "Ladies Night Out: A Typology of Women Patrons of a Male Strip Club", *Deviant Behavior*, 24: 333-52.
- Monto, Martin A. (2000) "Why Men Seek Out Prostitutes", in Ron Weitzer (Editor), *Sex for Sale: Prostitution, Pornography and the Sex Industry*. New York and London: Routledge.
- Monto, M. and Garcia, S. (2001) "Recidivism Among the Customers of Female Street Prostitutes: Do Intervention Programs Help?", *Western Criminology Review*, Volume 3, 2.
- O'Connell Davidson, Julia. (1998) *Prostitution, Power and Freedom*. Cambridge: Polity Press.

- O'Connell Davidson, Julia. (2002) "The Rights and Wrongs of Prostitution", *Hypatia*, 17, 2, 84-101.
- O'Connell Davidson, Julia and Jacqueline Sanchez Taylor. (2005) "Travel and Taboo: Heterosexual Sex Tourism to the Caribbean", in Elizabeth Bernstein and Laurie Schaffner (Editors), *Regulating Sex: The Politics of Intimacy and Identity*. New York and London: Routledge.
- O'Neill, Maggie. (2001) *Prostitution and Feminism: Towards a Politics of Feeling*. Cambridge: Polity Press.
- O'Neill, Maggie and Jane Pitcher. (2006) *Sex Work, Communities and Public Policy in the UK*. Research Paper. Cited with permission. www.safetysoapbox.com
- Padilla, Mark. (2007) *Caribbean Pleasure Industry: Tourism, Sexuality, and AIDS in the Dominican Republic*. Chicago: University of Chicago Press.
- Pateman, Carol. (1988) *The Sexual Contract*. Cambridge: Polity Press.
- Pheterson, Gail. (1993) "The Whore Stigma: Female Dishonor and Male Unworthiness", *Social Text*, 37, 39-65.
- Pitcher, Jane. (2006a) "Evaluating Community Safety Programmes and Community Engagement: The Role of Qualitative Methods and Collaborative Approaches to Policy Research", *Urban Policy and Research*, 24, 1, 67-82.
- Phoenix, J and S. Oerton. (2005) *Illicit and Illegal: Sex, Regulation and Social Control*. Cullompton, Devon: Willen Press.
- Prasad, M. (1999) "The Morality of Market Exchange: Love, Money and Contractual Justice", *Sociological Perspectives*, 42, 2, 181-215.
- Raymond, Janice G. (2005) "Sex Trafficking is Not Sex Work", *Conscience*, 26, 1.
- Research for Sex Work, No. 8, (2005) "Sex Work and Law Enforcement" June.
- Sánchez Taylor, Jacqueline. (2001) "Dollars are a Girl's Best Friend? Female Tourists' Sexual Behavior in the Caribbean", *Sociology: Identity Politics in the Workplace*, 35, 3.
- Sanders, Teela. (2002) "The Condom as a Psychological Barrier: Female Sex Workers and Emotion Management", *Feminism and Psychology*, 12, 561-66.
- Sanders, Teela. (2004) "The Risks of Street Prostitution: Punters, Police and Protesters", *Urban Studies*, 41, 9, 1703-1717.
- Sanders, Teela. (2005a) *Sex Work: A Risky Business*. Uffculme, UK: Willan Publishing.
- Sanders, Teela. (2005b) "It's Just Acting': Sex Workers' Strategies for Capitalizing on Sexuality", *Gender, Work and Organization*, 12, 4, July.
- Sassen, Saskia. (2002) "Global Cities and Survival Circuits", in Barbara Ehrenreich and Arlie Russell Hochschild (Editors), *Global Woman: Nannies, Maids and Sex Workers in the New Economy*. New York: Henry Holt and Company LLC.
- Schlosser, Eric. (2003) *Reefer Madness: Sex, Drugs and Cheap Labor in the American Black Market*. Boston and New York: Houghton Mifflin.
- Soothill, Keith and Teela Sanders. (2004), "Calling the Tune? Some Observations on Paying the Price: A Consultation Paper on Prostitution", *The Journal of Forensic Psychiatry and Psychology*, 15, 4, 642-659.
- TAMPEP. European Network for HIV/STI Prevention and Health Promotion among Migrant Sex Workers. (2004) *Final Report* No. 6.
- Truong, Thanh-Dam. (1990) *Sex, Money and Morality: Prostitution and Tourism in Southeast Asia*. London: Zed Books.
- UNESCO (2006) *Trafficking Statistics Database*. www.unescobkk.org/index.php?id=1022
- United Nations. (2000) *United Nations Protocol to Prevent, Suppress and Punish*

- Trafficking in Persons, Summary*. New York: UN.
- US Department of State. (2005) *Victims of Trafficking and Violence Protection Act of 2000: Trafficking in Persons Report*. www.state.gov/g/tip/rls/tiprpt/2005/
- Weitzer, Ronald. (2000) (Editor) *Sex for Sale: Prostitution, Pornography and the Sex Industry*. New York and London: Routledge.
- Weldon, Jo. (2006) "Show Me the Money", *Research for Sex Work*, 9.
- Willman-Navarro, Alys. (2006) "Making it at the Margins: The Criminalization of Nicaraguan Women's Labor under Structural Reform", *International Feminist Journal of Politics*, 8, 2.
- Willman-Navarro, Alys. (2008) "Safety First, Then Condoms: Commercial Sex, Risky Behavior and HIV/AIDS in Managua, Nicaragua", *Feminist Economics* (special issue on "AIDS, Sexuality and Economic Development").
- Wood, Elizabeth Anne. (2000) "Working in the Fantasy Factory. The Attention Hypothesis and the Enacting of Masculine Power in Strip Clubs", *Journal of Contemporary Ethnography*, 29, 1, 5–31
- ICPR. (International Committee for Prostitutes' Rights.) (1989) "World Charter for Prostitutes' Rights", in G. Pheterson (Editor), *A Vindication of the Rights of Whores*. Seattle & Amsterdam: Seal Press. pp.40-41.
- Wortley, S. and B. Fischer. (2002) *An Evaluation of the Toronto John School Diversion Program*. Toronto: Centre of Criminology, University of Toronto.
- supporting networks. Website includes an extensive list of organizations and resources. www.desireealliance.org
- International Committee on Sex Workers Rights in Europe. www.sexworkeurope.org/website
- Network of Sex Work Projects. (NSWP) Global group. *Research for Sex Work*. www.nswp.org
- TAMPEP European Network for HIV/STI Prevention and Health Promotion for Migrant Sex Workers. www.tampep.com
- Prostitutes' Education Network (PNET) www.bayswan.org/penet.html

Sex Trafficking Resources

- UNESCO (2008) *Trafficking Statistics Database*. www.unescobkk.org/index.php?id=1022
- UN Office on Drugs and Crime: (Various years) Trafficking and legislation reports. www.unodc.org/unodc/en/trafficking_human_beings.html
- Global Alliance Against Trafficking in Women (2008) A global network of non-governmental organizations addressing the core aspects of trafficking in persons: forced labour and services in all sectors of the formal and informal economy as well as the public and private organisation of work.. www.gaatw.net/
- Coalition Against Trafficking in Women (2008) *An international coalition working to end trafficking and sexual exploitation in all its forms*. www.catwinternational.org/

Alys Willman-Navarro
World Bank
Washington DC.
USA

AWillman@worldbank.org

Websites and Organizations

Sex Worker Rights Organizations

Desiree Alliance. Coalition of sex workers, health professionals, social scientists, professional sex educators, and their

Global Warming and Climate Change

Kristen A. Sheeran

Introduction

Climate change due to global warming poses an enormous challenge for policy makers. The environmental, economic, and social consequences of climate change are potentially devastating, the science surrounding the precise timing and magnitude of warming is still uncertain, the costs of mitigating climate change are significant, and the benefits to avoiding climate change are disparate across countries. Reducing global emissions of the greenhouse gases responsible for global warming will require international cooperation, but thus far, efforts to forge an international agreement to mitigate climate change have fallen short of their goal of preventing dangerous climate change. As the international community continues its negotiations over how to mitigate climate change, understanding the political economy of climate change takes on added importance. However, latterly some degree of broad agreement has been made at least about the severity of the problem (see HM Treasury 2006; Metz, Bosch et al 2007).

Global Climate Change

Climate change refers to a change in the Earth's climate in excess of natural climate variability that is attributed directly, or indirectly, to human-induced changes to the atmosphere's composition. Climate change is the result of global warming. Global warming is caused by increased atmospheric concentrations of greenhouse gases, the most significant of which is carbon dioxide. Atmospheric concentrations of carbon dioxide and other greenhouse gases trap a portion of the sun's outbound energy reflected by the Earth and raise the Earth's

surface temperature. This is the oft-referred to "greenhouse effect". In the pre-industrial era, carbon dioxide concentrations in the atmosphere warmed the Earth's surface to temperatures to which life on Earth had become adapted. The world's terrestrial ecosystems withdrew carbon from the atmosphere through photosynthesis and released carbon again through respiration and decay. The industrial revolution disrupted this global carbon cycle by dramatically increasing carbon emissions from the burning of fossil fuels for energy, transportation, and industry and land use change and deforestation. This led to excess accumulation of carbon and other greenhouse gases in the atmosphere and an intensification of the greenhouse effect.

Evidence of a warming trend is already apparent. Over the last 150 years, the Earth's average surface temperature has been rising. During the 20th century alone, average global surface temperature increased by .6 Celsius degrees (1 degree Fahrenheit). In the Northern Hemisphere, warming during the 20th century was more extensive than during any previous century. (IPCC 2001a). The 1990's proved to be the warmest decade, and 1998, 2001, and 2002 were three of the hottest years ever recorded, since scientists began tracking temperatures back in 1861. The current and projected rate of warming is cause for alarm, as it is without precedent during the last 10,000 years. Between 1990-2100, global average surface temperature is predicted to rise by an additional 1.4-5.8 Celsius degrees (2.5-10.4 degrees Fahrenheit). To put this rapid temperature change in context, recall that the Earth was only 5 Celsius degrees (9 Fahrenheit degrees) cooler during the last ice age (IPCC 2001a).

According to climate simulation models, small changes in the Earth's average surface temperature can lead to significant climatic changes (IPCC 2001a). Although many

uncertainties still remain about the exact magnitude and timing of climate change, scientists predict that, at a minimum, such changes will prove disruptive and exact a significant human and economic toll. Many of the predicted impacts of global warming are already evident. In the Arctic, air temperatures have increased ten-times faster during the last century than global mean-surface temperatures, leading to significant melting of the ice and permafrost layer that covers much of this region. Melting of the permafrost layer in Russia and Alaska has already caused significant structural damage. The Greenland ice-sheet appears to be melting at twice the rate than scientists had predicted, disrupting its ecosystem and the livelihoods of those who depend upon it. In the mid-to-high latitudes of the Northern Hemisphere, snow cover has already declined by roughly 10%. The annual duration of lake and river ice-cover shortened by two weeks during the 20th century. Significant glacial retreat is evident in all non-polar regions; Switzerland has already lost 66 percent of its glacier volume (UNFCC 2006).

Glacial melting, coupled with thermal expansion of seawater, is predicted to increase sea-levels, potentially displacing millions of people in low-lying areas and island states. Sea levels have already risen by 10 to 20 centimeters over pre-industrial averages. Sea levels may rise an additional 9 to 88 centimeters by 2100. Fresh-water infusion from glacial melting can damage fragile marine ecosystems and possibly disrupt ocean currents like the Gulf Stream that play a critical role in distributing heat on Earth (UNFCC 2006). Salt-water intrusion from rising sea-levels contaminates underground fresh-water supplies, contributing to the global shortage of fresh-water (UNFCC 2006).

Climate change will bring changes in precipitation patterns, leading to more severe

drought and desertification in arid and semi-arid regions and increased rainfall and flooding in others. Changed weather patterns will affect agricultural productivity, with declines in crop yields predicted to be the greatest in tropical and sub-tropical regions. Regions like Latin America and Africa can anticipate more acute food shortages and famines. Moreover, climate change is expected to increase both the frequency and severity of extreme weather events, such as monsoons and hurricanes. The Rhine floods of 1996 and 1997, the Chinese floods of 1998, the East European floods of 1998 and 2004, the heat wave in Western Europe that claimed more than 30,000 lives in 2003, monsoons that left 60% of Bangladesh underwater in 2004, and the increase in category five hurricanes in the Atlantic like Hurricane Katrina which wiped out much of the Louisiana and Mississippi Gulf Coast, are recent examples of weather extremes that many scientists attribute to climate change (UNFCC 2006).

Climate change threatens human health directly from weather catastrophes like floods and droughts and heat related illnesses and deaths, and indirectly through reductions in food and fresh water supplies and changes in the ranges of disease vectors such as dengue fever and malaria (IPCC 2001a). A recent World Health Organization study attributes more than 150,000 deaths and 5 million illnesses each year to climate change. The study finds that poor countries, particularly in sub-tropical and tropical regions, are especially vulnerable to losing lives to global warming, due to their climates and lack of infrastructure to combat public health care threats (McMichael et al. 2003).

Finally, climate change may disrupt the capacity of ecological systems to provide critical ecosystem services such as air and water filtration, protection against soil erosion, habitat, and carbon sequestration.

Not only will the loss of these ecosystem services impact the health and well-being of human populations, it may result in wide-scale extinction of plants and animals and loss of biodiversity (IPCC 2001a). Scientists predict that plant and animal ranges will shift towards the pole and higher in elevation, while plant flowering, bird and insect arrival, and breeding seasons may begin earlier in the Northern Hemisphere. If warming is extensive and rapid enough, plant and animal species may not be able to migrate quickly enough to find more suitable habitats, or they may find their migration blocked by human developments (IPCC 2001a).

Human Causes of Climate Change

Given the scientific evidence that the Earth's surface temperature is rising, and what climate models predict to be the likely consequences of such warming, preventative measures to avoid further climate change seem warranted. However, the Earth's climate has experienced periods of warming and cooling throughout its geological history. Herein lies the source of so much of the controversy surrounding climate change. To what extent can current global warming be attributed to human causes?

The Intergovernmental Panel on Climate Change (IPCC) is widely viewed as the authoritative scientific source of climate change information. The IPCC was jointly established by the World Meteorological Organization (WMO) and the United Nation's Environment Program (UNEP) in 1988 to review and assess the science, impacts, and economics of climate change. The IPCC issues regular assessment reports that reflect the global scientific consensus on climate change. These assessment reports forge the scientific basis for decision making under the United Nation's Framework Convention on Climate Change and for negotiations over the Kyoto Protocol, the

prevailing international climate change treaty. More than 2000 scientists worldwide contributed to the IPCC's third, and most recent, assessment report which was published in 2001.

According to the IPCC, strong new evidence now suggests that most of the observed warming over the past 50 years is due to increased emissions of greenhouse gases from human activities, not natural causes (IPCC 2001a). Climate models that simulate temperature changes from natural causes alone cannot account for most of the observed increase in average surface temperature over the last 50 years. Only when natural causes and anthropogenic causes are combined do these models explain a substantial portion of the temperature rise (IPCC 2001a).

Since the beginning of the industrial revolution in the mid-nineteenth century, atmospheric concentrations of greenhouse gases have risen dramatically, reaching their highest levels ever recorded in the 1990s. The atmospheric concentration of carbon dioxide has increased more than 30% since the pre-industrial era and is expected to double from its pre-industrial level by 2050. Carbon dioxide from burning fossil fuels is the largest single source of greenhouse gas emissions from human activities. Roughly 75% of the carbon emissions produced over the last 20 years is due to the combustion of fossil fuels for energy, transportation, and industry. The remaining 25% of carbon emissions is attributed to land use change, especially deforestation in the world's tropical regions. Forests play a critical role in the global carbon cycle, sequestering carbon emissions from the atmosphere and storing it long-term in vegetation and soils. The conversion of forest land to alternative uses releases much of the stored carbon into the atmosphere and prevents further sequestration. Despite the very high rates of deforestation that prevailed

throughout the 1990's, it remains likely that carbon sequestration worldwide exceeded carbon release from deforestation. However, if current rates of deforestation worldwide continue, the world's forests may no longer function as net carbon sinks (IPCC 2001a).

While human activities are responsible for increased emissions of other greenhouse gases such as methane, nitrous oxide, and carbon monoxide, carbon dioxide is by far the most important greenhouse gas. Not only do human activities produce more carbon emissions than other greenhouse gases, but carbon emissions in the atmosphere are much longer lived than other greenhouse gases. For these reasons, a reduction in carbon emissions from human activities remains the primary focus of climate change mitigation efforts.

Special Challenges for Policy Makers

Uncertainty, Inertia, Precautionary Principle

The overwhelming consensus of the international scientific community is that climate change is a real phenomena that is due, in part, to human causes. But does scientific evidence alone provide an imperative for reducing carbon emissions from human sources? Climate change poses some unique challenges to policy makers trying to decide this issue. Reducing carbon emissions from human sources will involve costs that policy makers will want to weigh against the projected benefits of emissions reduction. The dynamics of climate change and the uncertainty which still surrounds it, however, make it difficult to apply conventional cost-benefit analysis to the issue of mitigation. Firstly, predicting the exact scale and timing of climate change is difficult, given the complexity, inter-relatedness, and sensitivity of natural systems. Abrupt and potentially catastrophic changes to the Earth's physical systems remain possible as a result of climate change.

Some of the predicted changes, such as the melting of polar ice-sheets and changes in ocean circulation, will be irreversible. In general, the more extensive the warming, the more rapid its onset, and the less time human and natural systems have to respond to a changed climate, the more damaging climate change will be and the more costly adapting to climate change will prove.

Secondly, there is a degree of inertia that characterizes climate systems. Carbon dioxide is so long-lived in the atmosphere that stabilizing carbon emissions at their current levels will not stabilize carbon concentrations in the atmosphere. To stabilize carbon concentrations in the atmosphere, emissions need to decrease to a fraction of their current level. The lower the level policy makers stabilize carbon concentrations the sooner and more severe the decline in global carbon emissions. Even after stabilizing carbon concentrations in the atmosphere, surface temperatures will increase for a century or more, and the slow rate of ocean cooling will ensure that sea levels continue to rise for many centuries. Given this inertia, the damage caused by human interference in the climate system may not become readily apparent until it crosses certain as-yet-to-be identified thresholds, at which point, the damages may prove both devastating and irreversible (IPCC 2001a).

For these reasons, it is difficult to apply a standard cost-benefit assessment to the issue of climate change mitigation. Given the cumulative impact of carbon dioxide in the atmosphere and non-linear damages, a relatively small reduction in emissions may yield little appreciable benefit; in which case, the benefits of avoided damages seemingly do not justify the costs. However, if critical emissions thresholds are crossed, the damages of irreversible and potentially catastrophic climate change would almost certainly justify the costs of mitigation. Immediate and

significant emissions reductions will be necessary to avoid climate change, but the uncertain benefits of those reductions in terms of avoided damages may not manifest until some time in the future. If the worst consequences of climate change would be experienced one or two generations hence, the tendency to discount the future, to attach more weight to a dollar spent today than a dollar saved in the future, will bias cost-benefit analysis in the direction of not taking aggressive action in the present to reduce global carbon emissions. Add to this the reality that emissions reduction is a global public good that is most efficiently provided through international cooperation (see below), and it becomes clear that the decisions of individual countries to mitigate carbon emissions should not be based solely on a straightforward comparison of their own individual costs and benefits.

For these reasons, policy makers have increasingly looked to the precautionary principle to guide their decisions regarding climate change mitigation. The precautionary principle recommends taking preventative actions to avoid uncertain, yet potentially very costly future events. In 1997, more than 2600 economists, including eight Nobel prize winners, signed a statement in support of the precautionary principle that said, "As economists, we believe that global climate change carries with it significant environmental, economic, social, and geopolitical risks, and that preventive steps are justified". To implement the precautionary principle, scientists recommend that policy makers act now to reduce carbon emissions and build-in safety margins when establishing targets, strategies, and timetables for emissions reduction (IPCC 2001a).

Even if countries adopt a precautionary stance with regards to climate change, many issues still remain. Identifying the level of emissions reduction for each individual

country and minimizing the costs to countries of reducing their emissions are important issues confronting policy makers. Since reducing carbon emissions is a global public good, international cooperation is the only efficient way to reduce global carbon emissions. Yet, achieving international cooperation on these issues has proven the greatest obstacle thus far to global climate control efforts.

Climate Control as Global Public Good

Carbon emissions are transboundary pollutants that yield the same impact on global climate no matter where they are produced. Lowering emissions in any one country reduces the threat of climate change, thereby benefiting other countries as well. Climate control, because it is both non-rival and non-exclusive, is a classic example of a global public good. The problem with public goods is that there is insufficient incentive for individuals to provide public goods; therefore the level of public goods provided is sub-optimal. No one wants to pay to produce a public good that others benefit from at their expense. The rational strategy for any individual is to "free-ride" off of the public goods produced by others. An individual will provide a public good as long as the benefit to the individual from providing the public good outweighs the individual's cost of providing it. But because the public good generates external benefits for others, the individual who provides the public good captures only a fraction of the total benefit it creates. Society's demand for the public good is greater than the individual's and it exceeds the willingness of the individual to pay to supply it. The individual, therefore, will fail to provide the socially optimal level of the public good.

The implication of the public good problem for international climate control is clear: international cooperation is necessary

to ensure that the optimal level of global carbon abatement is achieved. Otherwise, countries will choose abatement levels on the basis of their own individual costs and benefits, ignoring the benefit that their abatement generates for others. By reducing emissions only to the point where a country's marginal private benefit is equal to its marginal private cost, countries produce too little abatement. The marginal social benefit of abatement exceeds the marginal social cost of abatement at this level. International cooperation is necessary to overcome the free-rider problem and to provide the optimal level of global carbon reduction.

Ideally, an international climate control treaty will satisfy two conditions. First, the treaty will achieve the optimal level of global emissions reduction. Second, the treaty will minimize the total cost of global carbon abatement. The cost of global abatement is minimized if the marginal cost of emissions abatement is the same for all countries. If marginal reduction costs differ across countries, shifting abatement from the country with the higher marginal cost to the country with the lower marginal cost would lower the total cost of global abatement.

An efficient treaty ensures that individual countries abate at levels that equalize their marginal abatement costs and that the sum total of individual countries' abatement is equal to the optimal level of global abatement. An inefficient treaty fails to satisfy one or both of these conditions. Three problems immediately arise for policy makers tasked with designing an efficient climate control treaty. Firstly, not all countries may benefit equally from climate control. Secondly, equalizing marginal abatement costs across countries may mean that some countries will abate much more than others. This opens up the possibility that some countries will benefit more from participating in an efficient treaty than others. Some

countries could be potentially worse off for their participation; for example, countries expecting only minimal damage from climate change and/or countries that can reduce emissions at relatively low cost (Sheeran 2006). For an international climate control agreement to be self-enforcing, all sovereign countries must find it in their own self-interests to voluntarily participate (Barrett 1994). Thirdly, it is also the case that countries are not equally responsible for the accumulation of greenhouse gases in the atmosphere that is driving climate change. Nor do all countries exhibit equal ability to pay for the costs of reducing greenhouse gas emissions. Satisfying these equity concerns may require countries to abate at levels that are not necessarily economically efficient.

These problems which confront policy makers in designing an efficient and equitable climate control treaty are not intractable, though their existence predictably complicates negotiations. As compared to the non-cooperative outcome where countries choose abatement levels independently of each other, an efficient climate control treaty will generate a greater net global benefit. This efficiency gain means that the treaty can make all countries better off. Whether or not an efficient climate control treaty actually makes every country better off, however, will depend upon how that efficiency gain is distributed. Mechanisms, such as emissions trading, that effectively separate out where emissions reduction takes place and who pays for it, and mechanisms, such as side-payments or exemptions, that redistribute part of the net global benefit for equity reasons or to ensure the participation of certain countries, exist and have been utilized to more or less success in international climate negotiations thus far. Examining the history of international climate change negotiations sheds important light on these challenges and

illuminates a path for future policy makers to take.

History of International Climate Control

UNFCC and Kyoto Protocol

Attempts to forge international cooperation on mitigating climate change began in 1992 with the United Nation's Framework Convention on Climate Change (UNFCC) in Rio de Janeiro, Brazil. At that time, the participating industrialized countries agreed to voluntarily reduce their greenhouse gas emissions. However, the need to strengthen the UNFCC soon became evident as most of the industrialized countries failed to meet their voluntary targets and emissions in some countries actually increased. More than 150 countries convened in Kyoto, Japan in 1997 to negotiate binding emissions quotas. The product of those negotiations, the Kyoto Protocol, remains the prevailing international agreement on combating climate change to-date.

The Kyoto Protocol establishes binding emissions limits for industrialized countries and a range of mechanisms to promote cost-effective compliance. Following the framework of the UNFCC, the Kyoto Protocol differentiates between two groups of countries worldwide, "Annex I" countries which are subject to emissions limits and "non-Annex I" countries which have no binding commitments. The list of "Annex I" countries includes 39 industrialized countries and countries with economies in transition. These countries account for 2/3 of global carbon emissions. The United States alone, accounting for less than 5% of the world's population, produces 25% of global carbon emissions. Although individual country commitments vary, Annex I countries, on average, are required to reduce emissions by 5.2% of their 1990 emissions levels during the commitment period 2008-2012. Table 1

details country commitments under the Kyoto Protocol.

Table 1. Country Commitment Under the Kyoto Protocol

Country	Target *
EU-15**, Bulgaria, Czech Republic, Estonia, Latvia, Liechtenstein, Lithuania, Monaco, Romania, Slovakia, Slovenia, Switzerland	-8%
United States***	-7%
Canada, Hungary, Japan, Poland	-6%
Croatia	-5%
New Zealand, Russian Federation, Ukraine	0%
Norway	+1%
Australia***	+8%
Iceland	+10%

Source: Adapted from UNFCC (2006)

*Target is defined as percentage change from 1990 emissions levels during the commitment period, 2008-2012.

** The 15 member states of the EU have reached their own agreement about how to distribute the 8% emissions target amongst themselves.

*** The U.S. and Australia have officially withdrawn from the Kyoto Protocol.

Policy makers confronted three challenges in designing the Kyoto Protocol. First, in keeping with the original UNFCC mandate, they were to establish emissions targets compatible with stabilizing greenhouse gas concentrations in the atmosphere at levels consistent with preventing catastrophic human-induced climate change. Secondly, to improve efficiency and increase the willingness of countries to participate, policy makers needed to design mechanisms to minimize the costs of meeting those targets. Lastly, to address historical imbalance in greenhouse gas production and differences in countries' abilities to pay for abatement, policy makers had to distribute the costs of preventing climate change equitably across countries.

The Kyoto Protocol provides countries with three mechanisms for minimizing the costs of meeting their emissions targets. First, the Kyoto Protocol allows for international emissions trading between Annex I countries. Emissions trading allows reductions to take place where marginal abatement costs are lowest. A country that reduces emissions by more than its required target can sell the “credits” for that reduction to other countries to use toward their own commitments, enabling those countries to avoid more costly reductions at home. Joint Implementation (JI) is the second cost-saving mechanism allowed under the Kyoto Protocol. Under JI, Annex I countries receive credit towards their own emissions targets for implementing projects aimed at reducing emissions in another Annex I country. Again, because marginal abatement costs differ across countries, the aim of JI is to lower total abatement costs by concentrating reductions in the Annex I countries where they cost least. Finally, the Kyoto Protocol’s third cost-saving mechanism, the Clean Development Mechanism (CDM), allows Annex I countries to receive emissions credit for financing projects that reduce emissions in developing countries. CDM projects can be unilateral, meaning that developing countries can undertake CDM projects in their own countries without an explicit Annex I partner and market the resulting emissions credits themselves.

Though emissions trading, Joint Implementation, and the Clean Development Mechanism may be effective in lowering global abatement costs, these mechanisms may benefit some countries more than others. Countries that can meet their emissions targets more easily than others may benefit from the sale of emissions credits, unlike other countries who may have to buy credits to meet their Kyoto commitments. In this respect, the assignment of emissions caps

under the Kyoto Protocol directly affects income distribution amongst participating countries. The Russian Federation, for example, is required to reduce emissions by 0% of its 1990 level as per the Kyoto Protocol. Emissions from industry in Russia declined somewhat unexpectedly throughout the 1990s, leaving Russia in the now enviable position of potentially having a surplus of carbon emissions credits to sell in the international market. Prior to withdrawing from the Kyoto Protocol, the U.S. had indicated its intent to meet most of its Kyoto commitment by purchasing emissions credits abroad. The demand of the U.S. for emissions credits could have driven up the price of emissions credits, precluding many smaller countries from the market. Developing countries, who are not required to reduce emissions under the Kyoto Protocol, can sell credits for emissions reduction they engage in as part of the CDM. In this respect, the CDM is a potential windfall gain to developing countries.

The Kyoto Protocol was designed to shift the burden of emissions reduction more heavily onto those countries that were most responsible for climate change and that were most able to afford emissions reduction. Concerns that these cost-saving mechanisms could allow countries like the U.S. to escape the burden of costly emissions reduction and forestall the transition away from fossil fuels prompted objections to full-scale use of these mechanisms. At present, the Kyoto Protocol does not specify a quantitative limit on the use of emissions trading or other cost-saving mechanisms to meet country commitments. It does, however, state that domestic actions must constitute a “significant element” of a country’s emissions reduction efforts (Kyoto Protocol 1997). Limiting the use of these mechanisms, however, comes at the expense of efficiency. An approach that allowed for full-scale use of these mechanisms, but that

required more extensive reductions from the countries whose costs were significantly lowered as a result, might have been preferable. Such an approach could have ensured that the global community shared in the efficiency gain generated by these mechanisms.

An additional complication in the negotiations over the Kyoto Protocol involved the role of carbon-sequestration by forests and other carbon sinks. The Kyoto Protocol requires Annex I countries to account “net changes in greenhouse gas emissions by sources and removals by sinks resulting from direct human-induced land-use change and forestry activities, limited to afforestation, reforestation and deforestation since 1990”, (Kyoto Protocol 1997). This means that Annex I countries can claim credit toward their emissions reduction targets for carbon sequestration within their own borders, provided that it has taken place since 1990. Effectively, Annex I countries will not have to reduce emissions from energy, transportation, or industry as significantly to meet their Kyoto targets. This mitigates the burden of complying with the Kyoto Protocol, especially for countries with abundant forest resources. Therefore, it’s no surprise that the crediting of sequestration activity by the Kyoto Protocol has always proved controversial.

At present, the Kyoto Protocol does not cap the total amount of credits countries may claim from sequestration within their own borders. However, it does limit the amount of sequestration credits countries can claim specifically from forest conservation. Policy makers were concerned that countries would claim credit for preserving forest areas that were not immediately threatened with deforestation where carbon sequestration would have continued unabated. Policy makers only wanted to grant Annex I countries credit for taking deliberate actions

to stave off deforestation or expand forest areas through reforestation or afforestation. Unable to credibly differentiate between forest areas where countries needed to take deliberate measures to prevent forest loss and areas that would have remained forested anyway, the Kyoto Protocol established a cap on the total amount of conservation credits countries could obtain. A cap wasn’t necessary for afforestation or reforestation, since these activities are almost always intentional.

A similar controversy surrounds crediting sink activities in developing countries under the CDM. Just as the CDM awards credits to Annex I countries for financing emissions reductions from energy, transportation, and industry in developing countries, the CDM should award credits to Annex I countries for financing the creation and expansion of carbon sinks in developing countries. In this way, the CDM would provide real incentives to developing countries to preserve and expand their forest areas. However, if Annex I countries can avoid reducing their own emissions by using the CDM to receive credit for carbon sequestration activities in the developing world, global emissions will actually increase *if* the carbon sequestration activity in the developing world they claim credit for is not additional to what would otherwise have occurred outside of the CDM. Identifying the baseline amount of sequestration activity that would have occurred in developing countries without the CDM is especially problematic, given that developing countries, unlike Annex I countries, are not required by the Kyoto Protocol to measure and account for net changes in emissions from sources or removals by sinks (Schlamadinger and Marland 2000).

For these reasons, the Kyoto Protocol limits sink activities under the CDM to only afforestation and reforestation, and limits the

amount of credits countries can obtain from such projects to no more than 1% of the country's base year emissions. This restriction limits the incentives the Kyoto Protocol provides developing countries to preserve and expand their forest areas. Because developing countries do not face mandatory emissions targets, they do not have the same incentives that Annex I countries do to limit emissions from deforestation or to increase sequestration by expanding forest areas. Given the global ecological and economic significance of preserving forests in the developing world, the inability of the Kyoto Protocol to provide developing countries with greater incentives for forest conservation is one of its most serious weaknesses.

Assessing the Kyoto Protocol

Equity concerns motivated a climate control framework in which countries were assigned differentiated responsibilities for reducing carbon emissions in recognition of their respective abilities to pay and their historic role in the build-up of greenhouse gases in the atmosphere. Developing countries were not required to reduce emissions, despite the projected increase in emissions from many of these countries in the near future. For example, in India, the world's fifth largest fossil-fuel emitting country, emissions increased 57% over 1990-1998. In China, the world's second largest fossil-fuel emitting country, emissions increased by 39% between 1990-1996 (Marland, Boden, and Andres 2000). Predicting the rate of emissions increase is more difficult for developing countries because emissions growth will be a function of the uncertain future growth rates of these economies and whatever initiative these countries take to reduce their emissions.

Perhaps the most serious weakness of the Kyoto Protocol is its inability to provide developing countries with direct incentives to

reduce their emissions. However, it is remarkable the extent to which developing countries are taking voluntary actions to reduce or slow the growth of their emissions. Examples include the phasing out of fossil fuel subsidies (e.g., Indonesia and China), establishing national goals for renewable energy use and energy efficiency (e.g., China, Mexico, India, Thailand, and the Philippines), and converting automobiles and public transport to natural gas (e.g., Argentina and India) (Biagini 2000). China, the world's most populous nation, actually reduced its greenhouse gas emissions by 19% between 1997 and 2000 while its economy grew by 15%. (Baumert and Kete 2002). According to Zhang (1999), China has adopted massive energy policy reforms over the last two decades aimed at increasing energy efficiency and conservation. China's sweeping measures represent emissions savings nearly equivalent to the entire U.S. transportation sector (Zhang 1999).

In comparison, emissions in the United States, the world's largest greenhouse gas emitter and third largest per-capita producer of greenhouse gas emissions, continue to rise. U.S. initiatives aimed at reducing greenhouse gases have been voluntary and not well-coordinated economy-wide (Baumert and Kete 2002). The recently much touted decrease in U.S. carbon-intensity, a measure of carbon emissions per dollar of GDP, is not will not reduce U.S. carbon emissions as long as U.S. GDP growth continues to rise. While more significant emissions reductions from developing countries will clearly be necessary over the next few decades, it is hard to expect developing countries to implement binding limits on their greenhouse gas emissions, when the wealthiest of the industrialized nations is unwilling to follow suit.

The U.S. dealt a potentially lethal blow to international climate change negotiations with

its decision to withdraw from the Kyoto Protocol. The Kyoto Protocol required ratification by at least 55 countries representing at least 55% of Annex I countries' emissions before it could enter into force. Without the world's largest emissions producer, implementation of the Kyoto Protocol required the participation of almost every other Annex I country. Ultimately, only Australia, Croatia, and Monaco joined the U.S. in refusing to ratify the Kyoto Protocol. The Kyoto Protocol entered into force in February of 2005, almost eight years after it had been initially written. It includes 161 states and regional economic integration organizations representing 62% of the emissions of the Annex I countries.

Will this level of emissions reduction be sufficient to prevent catastrophic climate change? According to most scientists and the IPCC, more significant emissions reductions will be necessary beyond the initial commitment period. If net emissions from human sources remain at their current level of 7 gigatons of carbon per year, atmospheric concentrations will climb to roughly twice their pre-industrial levels of 280 parts per million. To stabilize the atmospheric concentrations at their current levels would require an immediate 50-70% reduction in emissions and further reductions thereafter. To ensure that the atmospheric concentration remains below 550 parts per million, global annual average emissions could not exceed the current global average and would have to fall below the current average before the end of the twenty-first century. Global annual average emissions could be higher for stabilization levels of 750 to 100 parts per million, but achieving even these higher levels would require limiting the growth of global annual average emissions to no more than 50% above current levels on a per capita basis (IPCC 1995). Stabilizing atmospheric concentrations at less than 1000ppm would

only limit the increase in average global temperature to 3.5 Celsius degrees or less by the year 2100. In general, the lower the stabilization level, the smaller the increase in average surface temperatures, and the less costly the aggregate impact of climate change will prove to be (IPCC 2000). In light of such evidence, Kyoto's mandatory reductions seem woefully inadequate for achieving its goal of preventing climate change.

Climate Change Policy and the Future

The Kyoto Protocol is only a first-step in international climate control negotiations. Indeed, the next stage of international climate negotiations has already begun. The Parties to the Kyoto Protocol met in Montreal in December 2005 to plan for emissions reduction beyond Kyoto's first commitment period, 2008-2012. During the next commitment period, more extensive reductions, upwards of 30-60%, will have to be achieved. The participation of the United States, the world's largest carbon emitter, is absolutely critical in this regard. Getting a commitment from the U.S. to reduce its carbon emissions will likely prove the most difficult hurdle for international climate negotiations to overcome.

Other issues still remain. The current Kyoto Protocol limits the use of mechanisms such as emissions trading, Joint Implementation, and the Clean Development Mechanism. More extensive and effective use of these mechanisms, particularly as they relate to land use change and conservation, should be a goal of the next round of negotiations. Moreover, the next round should aim to provide more incentives for restructuring energy supplies away from fossil fuels and toward renewable and clean energy sources. Finally, the issue of developing country emissions must be dealt with. If equity warrants the continued exclusion of developing countries from

mandatory emissions limits, more technological transfers to the developing world will be necessary to both enable and encourage emissions mitigation in those countries. With the loss of life, morbidity, and economic cost predicted to be greatest in developing countries, international aid programs and development policies must also plan for significant climate-related damages.

Sadly, the focus of future climate negotiations must include adaptation. Increasing scientific evidence of rapid warming warrants prudent planning for the impacts of climate change. Scientists now project that the globally averaged surface temperature will warm by 1.4-5.8 Celsius degrees as compared to 1990 levels by 2100. Scientists predict that the globally averaged sea level will rise by .09 to .88 meters by 2100 (IPCC 2001b). Regional impacts of these changes will vary, but all regions will be negatively affected. Africa, Latin America, and Asia can expect an increase in floods and droughts, changes in seasonal river flows, diminished food security, a greater incidence of disease, and a decline in biodiversity. Small island states will be especially vulnerable to sea level rise and flooding, increased scarcity of fresh water supplies, and the loss of tourism. North America, Europe, New Zealand, and Australia can expect many of the same impacts from climate change. Weather-related disasters in North America and Europe are already on the rise and significant glacial melting has been observed. However, because these regions have greater adaptation capacities than the developing world, human and economic systems may be less affected. Indigenous communities which are dependent on climate-sensitive resources are the most vulnerable in these regions (IPCC 2001b).

It is in the polar regions where the most rapid, dramatic, and far-reaching effects of climate change may manifest. Climatic

changes in Polar regions could trigger centuries long, potentially irreversible climate changes, even after greenhouse gas concentrations are stabilized. Examples include large reductions in the Greenland and West Antarctic Ice Sheets, a significant slow-down in the ocean circulation of warm water to the North Atlantic that could radically alter Europe's climate, and warming-induced releases of terrestrial carbon from permafrost melting and methane from hydrates in coastal sediments that would magnify greenhouse gas concentrations and climate change. At this point, the likelihood of these events remains uncertain, but scientists expect the likelihood to increase with the magnitude, speed, and duration of climate change (IPCC 2001b).

Mitigating and adapting to climate change will remain major priorities for the global community throughout the twenty-first century. The threat of climate change, however, presents great challenges as well as opportunities. The other serious environmental problems confronting this generation - mass extinction, deforestation, fresh water scarcity, and desertification - are all global in scale. Lessons can and should be drawn from past and future international climate control efforts that can better inform future efforts aimed at preventing other global environmental disasters.

Selected References

- Barrett, Scott. (1994) "Self-Enforcing International Environmental Agreements." *Oxford Economic Papers*, 46, 878-894.
- Baumert, Kevin A. and Nancy Kete. (2002) *Will Developing Countries Carbon Emissions Swamp Global Emissions Reduction Efforts?*. World Resources Institute.
- Biagini, B. (2000) (Editor) *Confronting Climate Change: Economic Priorities and Climate Protection in Developing Nations*.

- Washington, DC: National Environmental Trust.
- Energy Information Administration (EIA). (1999) *International Energy Annual 1999*. Washington, DC: EIA.
- HM Treasury (UK). (2006) *The Stern Review on the Economics of Climate Change*. Cambridge, UK: Cambridge University Press. www.timcousins.com.au/stern_review_final_report.htm
- Intergovernmental Panel on Climate Change. (1995) *IPCC Second Assessment: Climate Change 1995. A Report of the Intergovernmental Panel on Climate Change*. New York: Cambridge University Press.
- Intergovernmental Panel on Climate Change. (2000) *Land Use, Land Use Change and Forestry: A Special Report of the IPCC*. New York: Cambridge University Press.
- Intergovernmental Panel on Climate Change. (2001a) *Climate Change 2001: Synthesis Report. Summary for Policy Makers*. www.ipcc.ch/
- Intergovernmental Panel on Climate Change. (2001b) *Climate Change 2001: Impacts, Adaptation, and Vulnerability*. www.ipcc.ch (See Metz et al 2006.)
- Marland, G.; T.A. Boden, and R.J. Andres. (2000) "Global, Regional, and National CO₂ Emissions", in *Trends: A Compendium of Data on Global Change*, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy.
- McMichael, A.J., D.H. Campbell-Lendrum; C.F. Corvalán; K.L. Ebi; et al. (2003) *Climate Change and Human Health*. World Health Organization.
- Metz, Bert; Peter Bosch; Rutu Dave; Ogun Davidson and Leo Mayer. (2007) (Editors) *Climate Change 2007—Mitigation*. New York: Cambridge University Press. 846pp. Contribution of Working Group III to the UN 4th Assessment Report of the Intergovernmental Panel on Climate Change. www.mnp.nl/ipcc
- Nordhaus, William D. and Joseph Boyer. (2000) *Warming the World: Economic Models of Global Warming*. MIT Press.
- Schlamadinger, Bernhard and Greg Marland. (2000) *Land Use and Global Climate Change: Forests, Land Management, and the Kyoto Protocol*. Washington DC: Pew Center on Global Climate Change.
- Sheeran, Kristen A. (2006) "Forest Conservation in the Philippines: A Cost Effective Approach to Mitigating Climate Change?", *Ecological Economics*, 58,2.
- Sheeran, Kristen A. (2004) "Equity and Efficiency in International Environmental Agreements: A Case Study of the Kyoto Protocol", in G.M. Muducumura and M.S. Haque (Editors), *The Handbook of Development Policy Studies*. New York: Marcel Dekker Publishing.
- Toman, Michael A. (2001) (Editor) *Climate Change Economics and Policy: An RFF Anthology*. Washington DC: Resources for the Future.
- United Nations Framework Convention on Climate Change. (2006) *Feeling the Heat*. Washington DC; UN.
- Weyent, John P. (1999) (Editor) *The Costs of the Kyoto Protocol: A Multi-model Evaluation*. Special Issue of *The Energy Journal*.
- Zhang, Z. (1999) "Is China Taking Actions to Limit Greenhouse Gas Emissions? Past Evidence and Future Prospects", in W.V. Reid and J. Goldemberg (Editors), *Promoting Development While Limiting Greenhouse Gas Emissions: Trends and Baselines*. New York: UNDP and WRI.

Kristen A. Sheeran
Economics for Equity and
the Environment Network, USA
ksheeran@e3network.org

Green Politics

Brian Chi-ang Lin

Introduction

Green politics is a lively mobilization of a variety of ideas, values, and reform proposals for promoting environmental conservation, civic engagement, peace, and social justice, and leading socioeconomic progress to a state of sustainable development. The emergence of green politics represents a fundamental deviation from the mainstream's persistent emphasis on economic growth and its negligence in environmental degradation and deficiency of a long-term vision for global sustainability (Young 1994). Green politics has been in gestation for many years in most western societies and has gained some success through organizing green activists and parties, especially in the European countries. The development of green politics such as the appearance of the Green Party has been seen as forward-looking movements for resolving deficiencies of electoral politics. The Green Party is not just another party to develop political partisanship, but its role is to rectify the emphases of traditional political parties on pursuing short-term economic and individual national interests. Thus, green politics has a strong appeal to its advocates as an alternative to parties of both the left and the right.

In principle, the development of green politics is ecologically based and shares a common concern for environmental justice and equity. The underlying green political ideologies, however, are very diverse and include, for example, eco-Marxism (Martin O'Connor 1995 and James O'Connor 1998), eco-feminism (Plumwood 2002), eco-socialism (Pepper 1993), and the greening of liberalism (Wissenburg 1998). The intellectual inquiry of green politics carrying the term green politics *per se* has been

voluminous in green literature (see, for example, Spretnak and Capra 1985; Rainbow 1993; Rüdiger 1995; Bomberg 1998; Barry 1999; Torgerson 1999; Radcliffe 2000; Toke 2000). The investigation of green politics, however, has continued to evolve and is not restricted only to studies carrying the term green (or environmental) politics. In recent years, political ecology has emerged as a burgeoning field for examining the political consequences of environmental change and the complex relationships between environment, politics, and society (Zimmerer and Bassett 2003 and Robbins 2004).

A well-protected natural environment is crucial for global sustainability and is considered an important type of international or global public good that can generate benefits that spill over borders, regions, and generations (Morrissey *et al.* 2002). To some extent, the majority of people living on earth are aware of the significance of protecting our natural environment. Nevertheless, compared with their enthusiasm for economic growth, their environmental concern is out of scale. Also, it is costly to exclude anyone from enjoying the environment even if they have not contributed to its sustainability, and this leads to free riders. The emergence of green politics, therefore, is in no way of purely political significance (for promoting green parties). More importantly, it could substantially enrich the public debate on pressing environmental issues (such as climate change and global warming), help us to better understand the relationship between economic growth and the environment, and help to advise the government to implement forward-looking environmental policies.

Green Thought, Values, and Protagonists

Green thought in economic literature can be traced far back to the concept of *the stationary state* developed by John Stuart Mill (1965:746-751). Winch (2004:111)

points out that Mill is one of the earliest green thinkers, whose “defense of a zero-growth society conveys the substance of his environmentalist concerns.” Mill’s virtuous zero-growth, stationary-state society is “a continuous state of dynamic equilibrium” in which all improvements in new technologies can be redirected towards a redistribution of wealth and the promotion of quality of life (Winch 2004:122). Lin (2006) also emphasizes that Mill’s *stationary state* is in line with contemporary analyses of a sustainable society and should be best understood as a society with unlimited growth in mental culture and improvements in economic equality.

From the perspective of practicing and promoting green thought and values, the early 1970s could be considered a major turning point. The green activists and groups during this time included the Earth Day movement, Greenpeace, and the early green parties formed in Australia, New Zealand, and the United Kingdom. The first Earth Day was organized by Denis Hayes and the late U.S. Senator Gaylord Nelson in 1970 to raise public awareness to environmental crises. To date, the Earth Day international network has reached more than 12,000 organizations in more than 170 countries, while the U.S. program has kept over 3,000 groups and over 100,000 educators coordinating countless community development and environmental protection activities. In 1971, a small group of anti-war activists set sail from Vancouver, Canada to take action against the U.S. nuclear testing on Amchitka Island, Alaska. They chartered an old fishing vessel the *Phyllis Cormack*, renamed her *Greenpeace*, and sailed off to the prohibited zone. Although the U.S. still detonated the bomb, nuclear testing on the island ended the same year, and Amchitka was later declared a bird sanctuary. Today, Greenpeace, based in Amsterdam, is one of the leading organizations for

promoting global environmental protection and has around 2.8 million supporters worldwide, with a presence in more than 40 countries.

On 23 March 1972, the world’s first-recognized green party, United Tasmania Group (UTG), was formed in Hobart, Australia. Two months later, in May 1972, the world’s first national-level green party, the Values Party, was launched at Victoria University of Wellington, New Zealand. Founded in Coventry in 1973 as the “People” or, as it was later known, the Ecology Party, the British green party is amongst the oldest in Europe. It can be clearly perceived that nearly all green activists have not merely questioned material consumption of modern societies, but also overtly addressed spiritual or metaphysical issues to steer people’s attitudes and values toward new social objectives. For instance, the Values Party’s 1972 manifesto claimed that “New Zealand’s peculiar malady is not physical poverty; it is spiritual poverty” (Rainbow 1993:25). The British Ecology Party stresses that the values of conventional politics are fundamentally flawed and has created the first edition of a *Manifesto for a Sustainable Society* (MfSS).

Two influential green writings which challenged mainstream economics and greatly inspired subsequent green activities also appeared in the early 1970s. In 1972, *The Limits to Growth* (LTG) report warned that our earth’s carrying capacity would be exceeded within 100 years if the present growth trends in population, pollution, production, and resource use persisted (Meadows et al. 1972). Then, in his well-remembered *Small is Beautiful*, E.F. Schumacher suggested a return to ecologically sound agricultural techniques and communal ownership for a better society, and challenged mainstream economics:

“Economists themselves, like most specialists, normally suffer from a kind of

metaphysical blindness, assuming that theirs is a science of absolute and invariable truths, without any presuppositions ... Buddhist economics must be very different from the economics of modern materialism, since the Buddhist sees the essence of civilisation not in a multiplication of wants but in the purification of human character ... A Buddhist economist would consider ... consumption merely a means to human well-being, the aim should be to obtain the maximum of well-being with the minimum of consumption" (Schumacher 1973:ch.4).

In Canada, the first green party to answer the call for an ecologically-oriented Canadian society, the unofficially named "Small Party" with a special reference to Schumacher's *Small is Beautiful*, was formed approximately one month before their 1980 federal election in the Maritimes.

In essence, green values are pluralistic and can be classified, for example, as *ecocentric* versus *anthropocentric*. Ecocentrism is concerned with the intrinsic value of non-human nature and, according to Pepper (1996), can be broadly divided into the following two types: mainstream Greens versus Green anarchists and Eco-feminists. Mainstream Greens advocate a simple lifestyle in a small-scale economy and acknowledge a positive role for government intervention, especially at the local level. Also, they prefer to adopt gradual, reformist methods to achieve their goals. Green anarchists and Eco-feminists, however, tend to use radical methods. They generally reject parliamentary democracy and favor non-hierarchical direct democracy. They also reject capitalism and favor common ownership of the means of production and income-sharing communes. From their perspective, the state has no role in their lives (Pepper, 1996: 43). Anthropocentrism recognizes humankind as "the only or

principal source of value and meaning in the world, and that nonhuman nature is there for no other purpose but to serve humankind" (Eckersley 1992: 51). The defining feature of the green moral theory, as Barry (1999: 27) argues, is "not the acceptance of ecocentrism but a critical attitude to anthropocentrism." This recognizes that humans are a differential within nature (not a separation from it), and that environmental ethics has to take the rights of other non-human species into account. Debates about green values and ethical issues have not yet ceased and remain vivid and healthy in green literature.

Green Parties and States

Generally speaking, the green social movement groups fall into the following three types: green parties; more radical environmental organizations such as national branches of Greenpeace and Friends of the Earth International (FoEI); and the green direct action groups such as Earth First! (Barry & Doherty 2001:591). On the European level, the first direct elections to the European Parliament (EP), the parliamentary body of the European Union, held in 1979, offered a timely flash point for many green activists to coordinate and promote their movements. Several green parties and lists such as the British Ecology Party, the French Europe Ecologie, Luxembourg's Alternative List-Resist (AL-WI), Belgium's Agalev and Ecolo, and Germany's SPV/Grünen participated in the 1979 election. Although none of them received a large enough percentage of the vote to win a European seat, they were successful in using this election to draw attention to ecological issues and recruit more members (Bomberg 1998: 84-85). After another decade of gestation and mobilization, the green political movement in European countries started in earnest in the 1980s.

In West Germany, the party *The Greens* (*Die Grünen*) was founded in 1979 and

officially organized on a national level in 1980. By advocating the four green pillars of ecological wisdom, social justice, grassroots democracy, and non-violence, the German Greens first entered the lower house of German parliament (*Bundestag*) with 27 seats in the 1983 election. In the 1987 national election, they increased their share of the vote to 8.3 percent with more than 40 seats in the *Bundestag*. Compared with other green parties in the world, the German Greens have been regarded as one of the most successful green parties. In the United Kingdom, the Green Party, previously named the Ecology Party, won two million votes and received a staggering 14.9 percent of the vote at the 1989 EP election. The British Greens, however, were prevented by their simple majority voting system from entering the EP. Overall, green parties across Europe achieved great success in the 1989 EP election by winning 28 seats in the European Parliament, a significant increase from 11 seats in the 1984 EP election.

Franklin and Rüdig (1995) have analyzed the 1989 EP election and identified three important determinants of green voting in the election. According to their study, environmental concerns appear to be the only common element to green voting across the European countries. In general, European green voters represent socially heterogeneous groups and are neither particularly left wing nor predominantly post-materialist. Thus, green voting primarily reflects an environmental concern, which may connect left-wing forces and post-materialist values to different degrees in different countries. The durability of green parties relies on their adaptability to external changing conditions and issues initiated by economic and social changes (Franklin and Rüdig 1995). Before February 2004, the European Green Parties appeared as a coalition known as the European Federation of Green Parties

(EFGP). On 20-24 February 2004, they decided to organize as a new pan-European party, the European Green Party, at the Fourth Congress of the EFGP held in Rome. In the sixth parliamentary term, the European Green Party has been linked with the European Free Alliance (known as Greens-EFA) by winning the above 40 seats of the MEPs in the election held in June 2004. Neumayer (2003) has empirically found that a strong green presence in the parliament has exerted a significant impact on the reduction of air pollution levels in many European countries.

Specifically, the development of the social environmental movement has been closely related to the state's stance. Hunold and Dryzek (2002) and Dryzek et al. (2003) use a fourfold classification with two dimensions to analyze the interaction of 'the state' and 'social movements' in a civic society. The *inclusive-exclusive* dimension measures the state's structural propensity for including or excluding special interest representation. An inclusive state is relatively open to a variety of interests and political actors, while an exclusive state limits representation to some political actors and denies access to others. The *active-passive* dimension reflects the state's propensity to affect or allow different interests. An active state is concerned with the social movements and organizations existing in society, while a passive state does little or nothing to undermine these particular movements and organizations and simply leaves them alone.

In this framework, Norway is an example of an *actively inclusive* state, where the environmental movement and concern have been comprehensively incorporated into the activities of the state. The U.S. represents an example of a *passively inclusive* state, where interest groups, including environmentalists, are allowed access to policy-making through lobbying activities. Inclusive states, such as

Norway and the U.S., tend to undermine democracy as a whole by draining civil society in the long run. After an oppositional group moves from a civil society to the state, the oppositional public sphere in question has gradually been exhausted and democratic authenticity has been attenuated. In this regard, The U.S. has exhibited less severe effects than in Norway.

The U.K. under Thatcher represents an example of an *actively exclusive* state, where the environmental movement was seen as a threat to market performance and was excluded by the state. Actively exclusive states inhibit diversity in the social movements and damage democratic qualities of a civil society. Under Thatcher's radical market liberalism, the U.K. created the peculiar phenomena of "minimal protest politics coexisting with relatively high membership of environmental groups and no access to the state" which had not appeared in other Western European countries (Dryzek *et al.* 2003:122).

Germany is an example of a *passively exclusive* state, where environmental groups which stood against the state were excluded from formal channels of policy influence. Passively exclusive states mostly ignore social movements and offer no points of entry for them. This may turn out in the long term to be beneficial for the democratic vitality of a civil society. The main reason is that democratization in the public sphere can be truly fostered beyond the state. Democratic authenticity in the public sphere in Germany, as characterized by its ecological modernization, associated sub-politics, and the strong connection between environmental values and state imperatives, ranks first among these four countries.

It might be further noted that the U.S. was the world's environmental policy leader in the 1970s. The U.S. established a national environmental protection agency (EPA) in

1970 and passed landmark legislation on clean air and water in subsequent years. However, unlike the green parties in other nations, the U.S. Greens have won elected office primarily at the local level. Nelson (2002) has found that green voting in the U.S. tends to be highly partisan. By using the adjusted League of Conservation Voters (LCV) scores for 1988-1998, he has noted that a senator's ideology is the most important variable for voting profiles on environmental issues. About 74 percent of measured ideology is explained by party affiliation and geographic location. Although Americans show strong support for environmental protection, their consensus declines when questions move from general environmental concerns to specific issues, and then frequently fail to send clear signals to policy-makers (Guber 2003). To Americans, they "prefer to buy green rather than vote green" (Guber 2003:156). The lack of green electoral success in the U.S. could be due to priority being placed on other issues (rather than environmental ones) during the hustle and bustle of election activities.

Trade and the Environment

Over the past decade, citizens living outside the U.S. have become more aware of the growing discrepancies in the attitude of the U.S. government toward trade and the environment. On the one hand, the U.S. government has shown aggressive leadership in promoting trade liberalization and free trade agreements. On the other hand, it has expressed persistent unwillingness to make an international commitment to environmental protection. The Kyoto Protocol came into force on 16 February 2005 with a total of more than 160 countries having signed the agreement. The two major countries currently opposed to the Kyoto agreement are the U.S. and Australia. Indeed, before the Kyoto Protocol was to be negotiated in December

1997, the U.S. Senate unanimously passed the Byrd-Hagel Resolution (S. Res. 98), sponsored by Democratic Party Senator Robert Byrd and Republican Party Senator Chuck Hagel, with a 95–0 vote in July 1997. The Resolution states that “the United States should not be a signatory to any protocol to, or other agreement regarding, the United Nations Framework Convention on Climate Change (UNFCCC) of 1992, at negotiations in Kyoto in December 1997, or thereafter, which ... would result in serious harm to the economy of the United States” (S. Res. 98).

To reduce the emissions of carbon dioxide and other greenhouse gases (GHGs), both the developed and developing countries ought to share common responsibility. Under the Clinton administration, the U.S. never officially evaded responsibility for its inordinate share of GHG emissions. Nevertheless, under the subsequent Bush administration, the U.S. directly sought to blame developing countries such as China and India, while eschewing America’s responsibility for their share of global environmental burdens (Harris 2004). In the 1970s, the U.S. was the global leader of the environmental movement and policy. Three decades later, the U.S. government and politicians have to some extent emerged as a potential obstacle to environmental progress in a global village. Many environmentalists outside the United States “bemoan the inability of America’s environmental movement to sway its own government” (Conca 2001:32).

In late November 1999, about 50,000 people gathered in Seattle to protest the World Trade Organization (WTO) talks. The goal of this impressive group action was not just to protest, but also to educate the public (Farley 2000). For many environmentalists, the values embodied by the WTO are not compatible in many ways with the green values and the WTO is merely an institution

created for catering to the interests of large corporations and investors (Weber 2001). The WTO, created in January 1995 as the dominant global trade regime, has continuously glossed over the disadvantages of free trade and the free market, their negative impact on the environment, the gap between rich and poor, the standards of trade unions, and so on. In fact, different people and communities should have wide latitude in choosing and building their own economic institutions catering to their specific interests, such as pursuing economic equality for social justice or developing a self-reliant type of economy with limited external trade (Lin 2006: 327). The U.S.-backed WTO and its associated agreements purport to accelerate the liberalization of trade and cannot meet the growing expectations of other concerns, such as environmental issues.

On 29 January 2000, in Montreal, over 130 countries concluded the Cartagena Protocol on Biosafety (CPB). The Biosafety Protocol was completed pursuant to the Convention on Biological Diversity (CBD) and meant to ensure the safe transfer, handling, and use of living modified organisms (LMOs). The Protocol entered into force on 11 September 2003 and nearly 140 countries have joined the agreement to date. Several major countries including the U.S. and Australia, however, have neither signed nor ratified the Protocol. The Biosafety Protocol is the first binding international agreement dealing with modern biotechnology, and has adopted the “precautionary principle”, a type of “safety first” approach to deal with scientific uncertainty. The EU countries greatly promote the precautionary principle as the European consumers generally remain skeptical toward the use of living or genetically modified organisms (GMOs) in food production (see, for example, Grunert et al. 2001; Magnusson and Hursti 2002; Saba

and Vassallo 2002). In addition to the potential danger to human health, GMOs also pose a threat to biodiversity since they artificially disturb the existing ecosystems and might lead to mutation, migration, and procreation (Gallagher & Werksman 2002:145).

Clearly, the relationship between the Biosafety Protocol and other international agreements, particularly the WTO, has been under potential tension. The Protocol Preamble, for example, emphasizes that “this Protocol shall not be interpreted as implying a change in the rights and obligations of a Party under any existing international agreements” and also concludes that “the above recital is not intended to subordinate this Protocol to other international agreements.” The Protocol specified the requirements for the LMOs being transported between exporters and importers. For instance, a Party in deciding the domestic use of LMOs is obligated to keep the Parties informed through the Biosafety Clearing-House (CPB, Article 11.1). This detailed report must also include a risk assessment. The requirements for LMOs used for food, feed, and processing (i.e., LMO-FFPs) are generally less strict but prospective importers for LMO-FFPs are obligated to make their laws and regulations available to the Clearing-House. They are also allowed to invoke the precautionary principle in deciding the importation of LMO-FFPs.

Winham (2003) points out that conflict between trade and environment regimes will likely expand out of the principle of ‘scientific risk assessment’ established in the WTO’s Sanitary and Phytosanitary Measures (SPS) Agreement and the ‘precautionary principle’ adopted by the Biosafety Protocol. The SPS Agreement is the WTO’s separate agreement specifying measures taken by member states to protect human and animal health (sanitary measures) and to protect plant

life or health (phytosanitary measures). It allows member states to set their own standards, but it also requires that “any sanitary or phytosanitary measure is applied only to the extent necessary to protect human, animal or plant life or health, is based on scientific principles and is not maintained without sufficient scientific evidence” (SPS Agreement, Article 2.2). Safrin (2002) argues that, under the rules of customary international law, the importance of the inclusion of the “savings clause” language in the Biosafety Protocol has been overestimated. The Biosafety Protocol represents an important achievement: its real significance lies more in what it says about the potential capacity for comity among nations in the face of a watershed technology than in what it says about its relationship to the WTO agreements (Safrin, 2002:628). It can be anticipated that the intense relationship between trade and the environment will be increasingly addressed in the arena of global environmental agreements and protocols.

Prospect of Green Politics

The new millennium will become an even more active era for worldwide environmental groups (Jacobs 1997). Environmental conflicts occurred in 1995 such as the disposal of obsolete oil rigs in the North Atlantic by Shell, French nuclear tests in the Pacific, and the mining of metallic ores in Papua New Guinea have shown the *global* nature of such conflicts (Low & Gleeson 1998). Although some progress has been achieved, as in the prevention of ozone layer depletion, the goal of reaching global sustainability has become even bleaker (Meadows et al. 2005), and environmental degradation has continued (see, e.g. Yi 2001; Diamond 2005). Environmental degradation has remained so difficult to resolve because it generally represents the serious consequences

of the simultaneous occurrence of market failure and government failure (Lin 2007).

Over the past several decades, green activists worldwide have participated in a variety of movements and shown unwavering commitment to our sustainable future. Despite fears that green parties might become more ephemeral in the future, it can be expected with great confidence that green activities will continue to evolve to take up the challenge. For instance, some have endeavored to raise our awareness of the rights of indigenous peoples and the preservation of their lands. In this regard, Darrell A. Posey, an influential figure for his advocacy of the rights of indigenous peoples, has emphasized that the development of traditional resource rights can protect the interests of indigenous peoples and strengthen the practice of their self-determination (Plenderleith 2004, ch. 14).

To resolve serious environmental problems and generate sustainability, the narrow academic circles must also be expanded to consider the significance of indigenous knowledge to advance the global knowledge commons (Dei *et al.* 2000). Indigenous knowledge is particularly abundant concerning the natural environment and has increasingly been recognized as an indispensable part of human capital and knowledge. Such knowledge is critical for the long-term development of human societies (see, e.g. Brokensha *et al* 1980; Fernando 2003). It can be foreseen that more and more people, after realizing the close relationship between indigenous knowledge and the natural habitats of indigenous peoples, will actively participate in promoting the rights and preserving the lands of these peoples.

Ideally, more and more academics can join the green movements to take their shared responsibility for the earth. The recent emergence of green economics (Lawson 2006; Wall 2006) and the development of

Post-Autistic Economics Network to promote pluralism in economics are a promising sign of such concerns and endeavors. Three decades ago, Nicholas Georgescu-Roegen, a pioneer in the field of ecological economics, advocated the abandonment of the two pillars of mainstream economics, discounting the future and maximizing utility, and pessimistically uttered that, with regard to future generations, “our policy toward natural resources must seek to *minimize regrets*” (Georgescu-Roegen 1977:375). The development of green politics is not just to help minimize regrets alone. More importantly, it awakes us to a sense of responsibility in striving for a sustainable future. In the years ahead, this growing sensed awareness will lead us to develop a more positive attitude toward urgent issues. Finally, let me conclude by briefly addressing a critical issue: water scarcities and shortages. Water sources are a type of abiotic resources and fresh water only accounts for less than three percent of the stock of water on earth (Daly & Farley 2004:87). The world is currently incurring a big shortage of water and some international legal arrangements might be further implemented for dealing with the impending crisis (Weiss *et al* 2005). A sustainable society cannot be maintained without sustainable water use. Undoubtedly, sustainable water use will soon be integrated into a central focus of green politics.

Acknowledgment: This article is devoted to the advancement of the Participatory Organization for a Sustainable Taiwan (POST) and to my students who have participated in green activities.

Selected References

Barry, John. (1999) *Rethinking Green Politics: Nature, Virtue and Progress*. London: Sage.

- Barry, John and Brian Doherty. (2001) "The Greens and Social Policy: Movements, Politics and Practice?", *Social Policy & Administration*, 35, 5, December, 587-607.
- Bomberg, Elizabeth. (1998) *Green Parties and Politics in the European Union*. London and New York: Routledge.
- Brokensha, D.; D. Warren and O. Werner. (1980) (Editors), *Indigenous Knowledge Systems and Development*. Lanham, MD:: University Press of America.
- Conca, Ken. (2001) "Green Politics in the Bush Era: Anti-environmentalism's Second Wave", *Dissent*, 48, 3, Summer, 29-33.
- Daly, Herman E. and Joshua C. Farley. (2004) *Ecological Economics: Principles and Applications*. Washington DC: Island Press.
- Dei, George J. Sefa; Budd L. Hall and Dorothy Goldin Rosenberg. (2000) (Editors), *Indigenous Knowledges in Global Contexts: Multiple Readings of Our World*. Toronto: University of Toronto Press.
- Diamond, Jared. (2005) *Collapse: How Societies Choose to Fail or Succeed*. New York: Viking.
- Dryzek, John S.; David Downes, Christian Hunold, David Schlosberg, with Hans-Kristian Hernes. (2003) *Green States and Social Movements: Environmentalism in the United States, United Kingdom, Germany, and Norway*. New York: Oxford University Press.
- Eckersley, Robyn. (1992) *Environmentalism and Political Theory: Toward an Ecocentric Approach*. Albany: State University of New York Press.
- Farley, Joshua. (2000) "Should Market Economists be Protesting the WTO too?" *Ecological Economics*, 33, 3, June, 337-340.
- Fernando, Jude L. (2003) "NGOs and Production of Indigenous Knowledge under the Condition of Postmodernity", *Annals of the American Academy of Political and Social Sciences*, 590, November, 54-72.
- Franklin, Mark N. and Wolfgang Rüdig. (1995) "On the Durability of Green Politics: Evidence from the 1989 European Election Study", *Comparative Political Studies*, 28, 3, October, 409-439.
- Gallagher, Kevin and Jacob Werksman. (2002) (Editors) *The Earthscan Reader on International Trade and Sustainable Development*. London: Earthscan.
- Georgescu-Roegen, Nicholas. (1977) "Inequality, Limits and Growth from a Bioeconomic Viewpoint", *Review of Social Economy*, 35, 3, December, 361-375.
- Grunert, Klaus G.; Liisa Lähteenmäki, Niels Asger Nielsen, Jacob B. Poulsen, Oydis Ueland, and Annika Astrom. (2001) "Consumer Perceptions of Food Products Involving Genetic Modification □ Results from a Qualitative Study in Four Nordic Countries", *Food Quality and Preference*, 12, 8, December, 527-542.
- Guber, Deborah Lynn. (2003) *The Grassroots of a Green Revolution: Polling America on the Environment*. Cambridge, MA.: MIT Press.
- Harris, Paul G. (2004) "International Development Assistance and Burden Sharing", in Norman J. Vig and Michael G. Faure (Editors), *Green Giants? Environmental Policies of the United States and the European Union*. New York: Cambridge University Press, 253-275.
- Hunold, Christian and John S. Dryzek. (2002) "Green Political Theory and the State: Context is Everything", *Global Environmental Politics*, 2, 3, August, 17-39.
- Jacobs, Michael. (1997) (Editor), *Greening the Millennium? The New Politics of the*

- Environment*. Oxford, UK: Blackwell Publishers.
- Lawson, Richard. (2006) "An Overview of Green Economics", *International Journal of Green Economics*, 1, Numbers 1/2, 23-36.
- Lin, Brian Chi-ang. (2006) "A Sustainable Perspective on the Knowledge Economy: A Critique of Austrian and Mainstream Views", *Ecological Economics*, 60, 1, November, 324-332.
- Lin, Brian Chi-ang. (2007) *More Government or Less Government? An Integrated View* Mimeo. Taiwan: National Chengchi University.
- Low, Nicholas and Brendan Gleeson. (1998) *Justice, Society and Nature: an Exploration of Political Ecology*. London and New York: Routledge.
- Magnusson, Maria K. and Ulla-Kaisa Koivisto Hursti. (2002) "Consumer Attitudes towards Genetically Modified Foods", *Appetite*, 39, 1, August, 9-24.
- Meadows, Donella H.; Dennis L. Meadows, Jorgen Randers, and William W. Behrens III. (1972) *The Limits to Growth*. New York: Universe Books.
- Meadows, Donella H.; Jorgen Randers and Dennis L. Meadows. (2005) *Limits to Growth: The 30-Year Update*. London: Earthscan.
- Mill, John Stuart. (1965) *Principles of Political Economy with Some of Their Applications to Social Philosophy*. Edited with an introduction by W.J. Ashley. First published 1848. New York: Augustus M. Kelley.
- Morrissey, Oliver; Dirk Willem te Velde and Adrian Hewitt. (2002) "Defining International Public Goods: Conceptual Issues", in Ferroni, Marco A. and Ashoka Mody (Editors), *International Public Goods: Incentives, Measurement, and Financing*. Boston: Kluwer Academic Publishers, 31-46.
- Nelson, Jon P. (2002) "Green Voting and Ideology: LCV Scores and Roll-call Voting in the U.S. Senate, 1988-1998", *Review of Economics and Statistics*, 84, 3, August, 518-529.
- Neumayer, Eric. (2003) "Are Left-wing Party Strength and Corporatism Good for the Environment? Evidence from Panel Analysis of Air Pollution in OECD Countries", *Ecological Economics*, 45, 2, June, 203-220.
- O'Connor, Martin. (1995) *Is Capitalism Sustainable? Political Economy and the Politics of Ecology*. New York: Guilford Press.
- O'Connor, James. (1998) *Natural Causes: Essays in Ecological Marxism*. New York: Guilford Press.
- Pepper, David. (1993) *Eco-Socialism: From Deep Ecology to Social Justice*. London and New York: Routledge.
- Pepper, David. (1996) *Modern Environmentalism: An Introduction*. London and New York: Routledge.
- Plenderleith, Kristina. (2004) (Editor) *Indigenous Knowledge and Ethics: A Darrell Posey Reader*. London and New York: Routledge.
- Plumwood, Val. (2002) *Environmental Culture: the Ecological Crisis of Reason*. London and New York: Routledge.
- Radcliffe, James. (2000) *Green Politics: Dictatorship or Democracy?* New York: Palgrave.
- Rainbow, Stephen. (1993) *Green Politics*. New York: Oxford University Press.
- Robbins, Paul. (2004) *Political Ecology: A Critical Introduction*. Malden, MA.: Blackwell Publishing.
- Rüdiger, Wolfgang. (1995) (Editor) *Green Politics Three*. Edinburgh: Edinburgh University Press.
- Saba, Anna and Marco Vassallo. (2002) "Consumer Attitudes toward the Use of Gene Technology in Tomato Production",

- Food Quality and Preference*, 13, 1, January, 13-21.
- Safrin, Sabrina. (2002) "Treaties in Collision? The Biosafety Protocol and the World Trade Organization Agreements", *American Journal of International Law*, 96, 3, July, 606-628.
- Schumacher, E.F. (1973) *Small Is Beautiful: Economics as if People Really Mattered*. London: Abacus.
- Spretnak, Charlene and Fritjof Capra. (1985) *Green Politics*. London: Paladin Grafton Books.
- Toke, Dave. (2000) *Green Politics and Neo-Liberalism*. New York: St. Martin's Press.
- Torgerson, Douglas. (1999) *The Promise of Green Politics: Environmentalism and the Public Sphere*. Durham, NC: Duke University Press.
- Wall, Derek. (2006) "Green Economics: An Introduction and Research Agenda", *International Journal of Green Economics*, 1, Numbers 1/2, 201-214.
- Weber, Martin. (2001) "Competing Political Visions: WTO Governance and Green Politics", *Global Environmental Politics*, 1, 3, August, 92-113.
- Weiss, Edith Brown; Laurence Boisson de Chazournes and Nathalie Bernasconi-Osterwalder. (2005) (Editors) *Fresh Water and International Economic Law*. Oxford & New York: Oxford University Press.
- Winch, Donald. (2004) "Thinking Green, Nineteenth-Century Style: John Stuart Mill and John Ruskin", in Mark Bevir and Frank Trentmann (Editors), *Markets in Historical Contexts: Ideas and Politics in the Modern World*. New York: Cambridge University Press, 105-128.
- Winham, Gilbert R. (2003) "International Regime Conflict in Trade and Environment: the Biosafety Protocol and the WTO", *World Trade Review*, 2, 2, July, 131-155.
- Wissenburg, Marcel. (1998) *Green Liberalism: The Free and the Green Society*. London: UCL Press.
- Yi, Zheng. (2001) *China's Ecological Winter* (in Chinese). Hong Kong: Mirror Books.
- Young, John. (1994) "Sustainable Development and Green Politics", in James E. Hickey and Linda A. Longmire (Editors), *The Environment: Global Problems, Local Solutions*. Westport, CT: Greenwood Press, 25-33.
- Zimmerer, Karl S. and Thomas J. Bassett. (2003) *Political Ecology: An Integrated Approach to Geography and Environment-Development Studies*. New York: The Guilford Press.
- Websites**
- Manifesto for a Sustainable Society. policy.greenparty.org.uk/mfss
- The Convention on Biological Diversity. www.biodiv.org
- Cartagena Protocol on Biosafety www.biodiv.org/biosafety
- Earth Day. www.earthday.net
- Earth First! Journal. www.earthfirstjournal.org
- Friends of the Earth International. www.foei.org
- Greenpeace International. www.greenpeace.org/international/
- Byrd-Hagel Resolution. www.nationalcenter.org/KyotoSenate.html
- Post-Autistic Economics Network. www.paecon.net.
- WTO's SPS Measures. www.wto.org/english/tratop_e/sps_e/sps_e.htm
- Brian Chi-ang Lin
Department of Public Finance
National Chengchi University, Taiwan
calin@nccu.edu.tw

Health Policy

Robert McMaster

Introduction

Health policy is in many respects a complex and contested terrain that reflects the wider undercurrents of social policy and social values. At the risk of being trite, health policy is a highly normative enterprise, inevitable perhaps since medicine is itself normatively grounded, and indeed the conception of care is also controversial and ethically laden (Fitzgerald 2004). This is perhaps brought into sharp relief by the *World Health Report* (2002), which emphasises the nature of global health disparities, principally in terms of infant mortality rates and life expectancies between high and low income regions. The Report states,

“Most of all [the findings of the report] emphasize the global gap between the haves and the have-nots by showing just how much of the world’s burden is the result of undernutrition among the poor and overnutrition among those who are better off, wherever they live. The contrast is shocking... [A]t the same time that there are 170 million children in poor countries who are underweight – and over three million die each year as a result – there are more than one billion adults worldwide who are overweight” (WHO 2002:8).

The Report highlights the dependence of individuals’ state of health upon poverty, behavioural patterns, culture, and wider social policies (Fine 2002; Galbraith 1973). Consequently, there are increasing calls for health policy to demonstrate a more holistic orientation (see, for instance, Hancock 1999, Mooney 2001, Sen 2002 *et al*) to reflect the underlying determinants of health. Hence, the strong advocacy that wider health policy *should* be tailored to the amelioration of poverty. Of course this is a reflection of an

underlying emphasis on the eradication of health inequalities (Sen 2002), and the notion that reasonable health is a basic human right (see Hancock 1999, WHO 2002).

Yet there is a perception of fiscal stress as witnessed by increasing expenditure on health care. This is attributed to the growth in demand for (and expectations of) medical services, partly reflected by the ageing population of developed countries; the adoption of new technologies that are not necessarily “cost effective”, and the relatively high rates of medical inflation (Newhouse 1993), prompting calls for greater efficiency in the provision of health care (especially from the nascent literature in mainstream health economics, see e.g. Diamond 1998). By contrast less developed regions, especially Africa and parts of southern Asia, are beset by relatively poor health care provision combined with growing epidemics in HIV/AIDS and poverty-related diseases (WHO 2002). Sometimes problems of under-provision are amplified by the costs of drugs.

Despite the contrasting nature of health problems, generally between wealthy and poor regions, health policy has been similar in that it has embodied an accent on reform of the structure of provision, and finance: what Iriart *et al* (2001) term the *transnationalization* of health policy. There is also an emphasis on *curative as opposed to preventive services*. This paper critically reviews the nature of health policy in its broadest terms. The rationale for policy reform is outlined. Comparative examples and the roles of international agencies is then examined. Lastly, health policy patterns are critically analysed.

Health Care Policy Reform:

The Economic Rationale

The most prominent aspect of the reform of health care provision relates to the structure of health care delivery. Structural, or

governance, change is motivated to varying degrees by an increasing recourse to an efficiency rubric. Mainstream economic approaches, such as new institutional economics, public choice models of bureaucracy as well as health economics, have to some extent displaced earlier neoclassical, Pigovian notions of market failure. Instead new institutionalism and public choice highlight state failure, which essentially revolves around the absence, or attenuated influence, of efficiency aligned incentives. Moreover, the seemingly inexorable growth of the state, and state expenditures (see Baumol 1993) is viewed as a potential drag on economic growth. The principal policy implications suggest vertical disintegration of state activities and a ‘contractualisation’ of welfare provision as vehicles for enhancing the efficient provision of state services (for example, see Niskanen 1968, Inman 1987, Williamson 2000), and enacted during the 1980s notably in the UK and USA.

The development of such agency and information-theoretic economics has been influential in tailoring the suite of policies advocated and, to varying degrees, enforced by international agencies, such as the International Monetary Fund (IMF), the World Bank (WB), the World Trade Organization (WTO) (see e.g. Fine 2001; Price et al 1999; Stiglitz 1998), and even WHO (Keaney 2002) since the 1980s especially in low-income regions. Much of this impetus comes from the USA (Rice 1998), and is a manifestation of more general change in the systems of welfare provision (for example, Fine 2002 and Reich 2002). This has prompted some debate about the appropriate role of the private sector in health care provision. Indeed, Le Grand (2003) in his foreword to a recent WB publication highlighting the “potential” of private participation in health care provision, perhaps

hits the nail on the head of the rationale underpinning the economics of this when he observes:

“Quite why it should be morally objectionable to make profits from the provision of health care than in other areas of equal or even greater importance to human welfare where private provision was common, such as food or housing, was never made clear”.

Concisely, from a mainstream economic theoretical perspective, tempering informational asymmetries favouring clinicians (or other state professionals), through the invocation of contractual and specific rivalrous relationships between health care providers, is considered to adjust agents’ incentives in a cost reducing fashion (see Enthoven 1994). Physicians, in particular, are frequently presumed to either be unaware of the cost implications of their actions; or to actively engage in inefficient activities to elicit utility gains: some mainstream health economists arguing that this arrangement is tantamount to a form of moral hazard (see for example, Pauly 1986). Indeed, many medical procedures and technological advances are often claimed *not* to be cost effective, yet enhance medical capabilities (Newhouse 1993). This supply-side ‘failure’ is compounded by demand-side inefficiencies: frequently patients do not directly bear the costs of treatment, and have no incentive to ration consumption. In effect, this incentive arrangement implies both over-provision and over-consumption of health care. Mainstream health economists recognise that patients tend to be passive in articulating demand; instead relying on clinical advice and guidance—the classic principal-agent scenario. Hence, reform should be concentrated on the supply-side, and is of a pattern that should furnish agents with incentives to pro-actively manage limited resources among different health care needs

in an efficient manner, and increase agents' accountability.

Institutionally, market-oriented reform further encapsulates the adoption of "new public management" (see for example, Light 2001). This term has been used to refer to managerial practises that typically include an emphasis on outcomes or (certain measurable) outputs as opposed to inputs, the vertical separation of organizational units into provider and client roles accompanied by some devolved budgetary responsibilities, thereby associating different parts of an organization through more contractual processes, and contracting out functions as a means of enhancing 'consumer' or 'client' options, and generating efficiency.

Some mainstream health economists recognising "consumer cost illusion" associated with third party payment arrangements for medical services have further advocated "demand-side" market-oriented reform in the embodiment of patient cost-sharing, i.e., the increased employment of user charges (see Rice 1998 for a critical discussion).

Nature of Health Care Reform: Some Comparative Examples

This section briefly outlines and contrasts health reform in selected countries: China, Germany, New Zealand, Netherlands, UK and USA. These countries were selected on the basis of demonstrating contrastive aspects of a general pattern of an increasing market-orientation in health care provision.

Arguably it is possible to identify such a broad pattern throughout Europe, where the state's role has shifted from direct provision to a more outcomes-focussed contractarian approach (see for example, Saltman 2002). Most notably in 1990 the UK initiated what was injudiciously termed as an "internal market" within its national health service (NHS). Principal amongst the organisational

changes were the separation of component bodies into purchaser and provider roles accompanied by devolved budgeting. In this way it was envisaged that purchasers (primary care providers, mainly general practitioner fundholders) would act as gatekeepers on expanding demands on the health service and their ability to switch between potential suppliers of secondary care (hospitals organised as NHS trusts) would further encourage efficiency-enhancing behaviour in those latter bodies. In effect the reform attempted a seeming contradiction: on one level to mimic what was taken to be the model of efficient market governance (and hence devolved responsibilities), and at another strengthen the regulatory role of central government (clear centralization).

The incoming "new" Labour government in 1997 initiated a further set of reforms that it claimed would "abandon" the "internal" market. Far from "abandoning" the 1990 model these subsequent reforms have retained its centrepiece: the purchaser-provider split (see for example, Light 2001, and McMaster 2002). Indeed, the initial phase of reform concentrated on horizontally integrating health care providers into larger bodies, while maintaining the vertical separation between primary and secondary care bodies. Moreover, greater devolution of financial responsibilities accompanied restructuring. Primary care groups/trusts "commissioned" referrals to secondary care providers, maintaining the structure of the 1990 budgetary arrangements—primary care providers acting as gatekeepers in articulating demand, and referrals being funded on a capitation basis. In saying this, however, the government is establishing a national tariff structure for commissioned activities in order to mitigate price fluctuations that may be a consequence of price competition.

Subsequent legislation has sought to increase the reliability of information and

establish a formal performance indicators, promote public-private partnerships and offer opportunities for increased financial autonomy. Smith (2002) observes that performance management explicitly relies on quantitative results and methods. Not only does this reform package resonate with “new public management” by focussing on what are deemed to be measurable outcomes, it also seeks a greater involvement from the private sector in terms of the provision and maintenance of capital assets, and the incorporation of private sector management techniques. Specifically, all UK state bodies are compelled to consider employing the private sector in financing, designing, building, operating and maintaining new capital assets, such as new-build hospitals. The private sector retains ownership of such assets, effectively leasing them to NHS authorities. In terms of endeavouring to incorporate what are taken to be private sector techniques, the 2000 NHS Plan states,

“The NHS currently lacks the *incentives* many *private* sector organisations have to improve performance, ...” (Department of Health 2000:28; emphasis added).

Consequently, there is a belief that restructuring will serve to hone incentives. The latest element of restructuring tied-in with sustained “excellence” as measured by performance indicators is to be the establishment of foundation hospitals. Again, as with primary care, foundation hospitals are to have extensive financial and managerial autonomy: potentially extending to some discretion over the range of services offered, some terms of employment, and the right to invest any financial surplus.

The Dutch and German social insurance systems demonstrate extensive similarities (Schut et al 2003), although importantly, in the Dutch system individuals earning more than a specified threshold income are no

longer entitled to social insurance; they must purchase private insurance coverage.

Unlike the UK, the Dutch health care system is a hybrid between public, private and non-profit providers and for-profit insurers, although the state is responsible for the accessible provision and quality of health care (Grit & Dolfsma 2002). Throughout the 1990s health care providers were de-regulated in an attempt to encourage greater competition, and extend the profit motive. Insurance was overhauled from a system of retrospective reimbursement to prospective payment: the objective being to adjust incentives to such that insurance enrolees had a stake in “efficient” health care provision. Insurance funds were also encouraged to spatially compete on the basis of price.

By contrast in Germany the scope for price competition among insurance providers or funds is restricted as these bodies jointly negotiate with health care providers. Hence, the basis for price competition resides in reducing costs and/or opting for favourable risks. The German system of social insurance was liberalised in the 1990s to create “socially bounded competition” (Schut et al. 2003) to encourage greater efficiency. Yet efficiency was not the only metric: insurance reform had the aim of improving access for all potential insurees, and to equalise contributions across the population. As with the Netherlands, German health care reform has not been as radical as the UK in its compass of restructuring health care delivery. New Zealand is the closest to the UK in terms of the scale and nature of reform.

New Zealand, like the UK, has a national health service funded and provided by the state. In the 1990s New Zealand embarked upon a programme of vertical disintegration that attempted to inject rivalry between providers, regional health authority contractors and the private sector. This was accompanied by the expansion of user fees,

especially for secondary care. Aspects of the initial 1991 programme have either been amended (as in restructuring) or retrenched (as in the expansion of user fees to selected secondary care procedures), although primary care fees have increased (Flood 2003, provides a detailed account). The emphasis on contracting and restructuring is similar to the UK's on-going reforms, and has certain complementarities with Enthoven's (1994) model of "managed competition" in that there was intent to introduce competition in finance. Notwithstanding some retreat from market-orientation the management of provision is heavily influenced by a financially driven efficiency ethos in providing an efficient level of care (Fitzgerald 2004).

The foregoing indicates that in different health care systems the concern with "efficiency" has accompanied the growth in health care expenditures, and the popularity of wider neo-liberal policies (Fine, 2001). Yet there have been important differences in the pace of adopting such policies. The UK and New Zealand have been markedly radical in the reconfiguration of provision. *Prima facie* the UK has not opted to radically alter overall financial arrangements in that competition between insurers has been eschewed; yet restructuring has involved distinct changes in the manner that NHS bodies are financed.

US health policy is not notable for reform as such, as for the failure of reform proposals. A decade ago the Clinton administration ventured to universalise health insurance coverage through a model of managed competition. The causes of the failure of the Clinton reform are beyond the parameters of this chapter (e.g. see Flood 2003). Failure has left a legacy of considerable inequalities in insurance coverage with estimates that well over 40 million Americans have either no insurance or are under-insured (Davis, 2000), although states have to varying degrees

attempted to increase accessibility to care; the system is highly fragmented. Further, there has been some recourse to the use of user charges, even in Medicare and Medicaid the federally provided facilities for the aged and poor respectively. As Flood (2003) observes this is a highly regressive finance mechanism, imposing the greatest burden on the financially more vulnerable. The increased utilisation of user charges reflects widely held concerns about the levels of health care expenditure in the US. Despite its reliance on a market system for both delivery and finance functions many commentators are paradoxically concerned about the efficiency of this arrangement (Nichols *et al.* 2004). Indeed, the US allocates substantially more resources to prolonging the life of the (wealthy) elderly than it does to preventive and primary care services. Relative to other industrialised economies it has a comparatively high infant mortality rate and low birth-weight babies, suggesting considerable inequality (cf. Deaton & Lubotsky 2003). Yet it seems that egalitarian pressures are muted in contrast to the quest for technological advances in care for those who can afford it.

Surprisingly, this may be a concern following radical reforms of the Chinese system. Liu and Mills (2002) report that there has been a growth in the employment of user charges accompanied by a reduction in the level of state budgetary support for public health services. Yet the scale of China's reform is astonishing as it impacted on each level of the country's public health institutes. Succinctly, throughout the 1980s and 90s the Chinese government combined a policy of budgetary retrenchment with liberalisation of rivalry between health care facilities. These bodies were to implement user fees to substitute for the withdrawal of state funding: state financial support is designed only to meet salary costs, user charges are intended to

cover other costs (Khaleghian & Das Gupta 2004). Contentiously, from a mainstream economic perspective the reform is laudable: it should encourage increased productivity and efficiency in provision as agents' incentives adjust to a competitive imperative, especially since health bodies have a right to retain any financial surplus and distribute it among their staff. Costs should fall following rivalry (health bodies have to accommodate any financial losses), consumer choice should be improved, and the accountability of agents will also be enhanced. In effect, "bureaucratic" incentives have given way to, what is considered to be, "entrepreneurial" incentives (cf. Saltman 2002).

Liu and Mills (2002) conclude that while the Chinese reforms may have improved the accountability of individual public health institutes, there have been some highly controversial, albeit unintended, consequences. The introduction of user charges has led to a decline in the provision of preventive services as government expenditure has become more concentrated on curative activities, and health education. There has been a discernible increase in the incidence of some communicative diseases during the period of the reform. Moreover, the authors also detected a shift in the range of activities, undertaken by health bodies, towards those offering revenue potential. While preventive activities declined there was an increase in fee-based health inspections, which led Liu and Mills (2002:1697) to observe:

"[T]he revenue-oriented behaviour of providers had led to the provision of unnecessary inspections and services for those who can pay, and the under-provision of necessary services for those who are not able to pay. It is evident that market-oriented financing reforms of public health services should not be considered a policy option".

Again, the common theme emerging from the foregoing contrastive outline is the increasing concern with efficiency and incentives. Indeed, even in circumstances where equity is prioritised, especially in health policy pronouncements concerning low-income countries, the narrative on appropriate "supply-side" incentives is prominent (see e.g. Brinkerhoff 2003).

Supra-National Influences on Health Policy

Health policy in low-income countries is subject to potentially considerable external influence. As noted, WHO, IMF, WB, and WTO all exert some influence on the nature of health policy. For instance, WHO has been instrumental in endeavouring to develop a global manifesto. Under the auspices of WHO, Sachs, *et al* (2001) advocated the implementation of the United Nations' laudable millennium development goals: reduction in poverty and "marked improvements" in the health of the poor. The report recognises the importance of preventive actions as a means of reducing the incidence of cardiovascular diseases, diabetes and cancers. It seeks the establishment of a global fund to "fight" AIDS, tuberculosis and malaria in low-income regions, particularly sub-Saharan Africa, where problems are exceptionally acute).

Sachs, *et al.* also correlate improved health conditions with economic development and growth. They appeal to high income countries to bridge the "funding gap" in low income countries' potential to meet what it terms as "essential intervention costs"—calculated at approximately \$34 per capita per annum. "At most", the funding gap is estimated at \$19 per capita per annum (or estimated at 0.1% of donor GNP), with the lowest income countries able to contribute \$15 per capita per annum (by 2007) *if* they adhere to a "sound" macroeconomic

framework and establish quantifiable operational targets.

Sachs, *et al.* assert that each country should define any programme of “essential interventions” on the basis of four criteria:

1. The programme is technically efficacious, i.e., it will be effectively delivered.
2. Targeted diseases should impose a “heavy burden” on society.
3. The social benefits of the programme should outweigh the social costs of intervention.
4. The needs of the poor.

This resonates with the WB’s emphasis on the benefits of the development of measurement techniques and human capital, where there is an association between health status and economic growth and “social capital”, in addition to the bank’s strong promotion of trade liberalisation. Other international bodies, such as the IMF and WTO, buttress this. The WB as part of its objective of tackling poverty through improvements in health status and health care provision in lower income regions has advocated an extensive role for the private sector in health care supply. As the bank frequently cites, the private sector is often the main medical care provider in lower income countries, especially of primary and curative care. It recognises that health care providers are regularly unqualified, and/or “traditional”, and/or pharmacists, but it stresses the need for the continuance of private delivery enabled by either direct or indirect state finance (Harding & Preker 2003). There is a tendency to adopt criteria that invoke governments employing “the *least* intrusive instrument that will achieve the desired objective” (Harding & Preker 2003:14; emphasis added). Although not necessarily reflecting the views of directors of the WB (or its member states), Harding and Preker’s guide to private

participation in health services is indicative of the broad thrust of policy. Thus, in their spectrum of intrusiveness direct provision is viewed as the most intrusive, followed by financing, then regulation and mandates, and information as the least. Accordingly, and drawing on concepts such as externalities, public and merit goods, direct provision may be apposite for rural hospitals and clinics, preventive services, such as immunizations, and sanitation. By this token a contractarian approach is supported “where appropriate”.

Nevertheless, Iriart *et al.* (2001) and Lloyd-Sherlock (2000) report that in Latin America the WB has tended to “impose” a single contractarian model, characterised by the separation of finance and delivery and the introduction of private enterprises in both, despite historical, social and epidemiological differences between countries. For instance, Iriart, *et al.* (2001) highlight how Argentinean and Brazilian reform proposals are derived from the USA. This involves the “..., subordination of health professionals to an administrative-financial logic” (Iriart *et al.* 2001:1245), and a reduction in independent clinical practices, as physicians become tied to insurance corporations and/or larger clinical centres. This model of managed care has engendered concerns that the orientation of management is on investment returns as opposed to restricted access for vulnerable groups (Iriart *et al.* 2001; Lloyd-Sherlock 2000).

A more contractarian orientation and increased private involvement in the provision of medical care enhance the potential for international trade through the general agreement on trade in services (GATS): a further manifestation of “transnationalization” of health policy. GATS, as administered by the WTO, relates to trade in commercial activities, and although this precludes those state activities deemed not to be commercial, such as

hospital services, from international competition (WTO, www.wto.org/), there is considerable latitude in other health care related areas, especially pharmaceuticals. The highest profile and most controversial aspect of trade in drugs arose recently with certain low and middle-income countries' threats to duplicate particular palliative drugs, which mitigate AIDS-related conditions, that were priced at an unaffordable level for the vast bulk of such countries' populations. This represented a clear breach of copyright, and as such corporations argued that this threatened innovative practice in the whole industry. DiMasi et al. (2003) observe that there are extensive sunk costs involved in the development of drugs, with only a relatively small chance of approval for use (just over one-fifth of those drugs that commence with human trials are eventually approved). After negotiations the corporations offered price discounts and donations, a gesture lauded by Sachs et al. (2001) who argued for further actions of this nature.

WHO, the WB (and the IMF), and WTO all influence health policy implementation towards what they frequently terms as a "mixed" provision, with opportunities for a contractarian orientation to be encouraged. There is also a stress on the adoption of quantifiable target-setting regarding outcomes, which is accompanied by an emphasis on the dissemination of "best practice". Yet there is considerable controversy surrounding the trajectory of health policy.

Critical Analysis

The increased market orientation of many health care systems, although not all (see for instance, Forget's (2002) and Sullivan and Mustard's (2001) accounts of Canadian "resistance" to US influenced reform) has been subject to extensive criticism (see, e.g. Keaney 2002; Light 2001; Rice et al 1998).

Notably, many of these assessments are based on criticisms of mainstream economics and health economics (Rice 1998). Although Saltman (2002) maintains that economists have not increased their policy prominence; it may nonetheless be contended that a certain type of economics has been highly influential in health policy formation. Fine (2001:143), for instance, describes the economic model employed as an analytical frame in the "Post Washington Consensus" (between the IMF and WB) as a "paler version of (the) earlier Keynesian - welfarism - modernisation perspective". The consensus is grounded on the approaches outlined above, which are reductionist in that they incorporate rational choice (instrumental rationality) as agents' sole motivating force, and is context independent in that it relies on the application of generalised principles. Moreover, "the market" is generally considered to be ubiquitous (cf. Williamson 2000). Reference to the health economics section on the WB website, reveals the embeddedness of the WB's analytical framework within mainstream economics.

As Fine (2001) and Keaney (2002) et al. convincingly argue, the presumption of a ubiquity of markets combined with the notion that health is somehow akin to a capital stock; a component element of human capital, supports the commodification of health. The mainstream health economics literature presumes this with conceptual analogies to individuals rationally trading-off their stock of health in order to attain utility maximisation. By logical deduction health care provision must be a commodity, which entails trading, assuming assigned property rights, at given prices. This contractarianism underpins the principal-agent approach, and reduces all human relations essentially to those of exchange without recognising the importance of institutions, beyond constraints, to human interactions. Hence the

mantra that through market orientation state failures, such as corruption or the inappropriateness of “bureaucratic” incentives, will be addressed. Such is the basis of Harding and Preker’s approach. Nonetheless, if low-income countries are considered to have weak and corrupt states market oriented reform cannot remedy this. The market as an institution does not reside in splendid isolation, but is a part of a complex system of institutions. Markets *need* strong states to function (see, for example, Hodgson 1999).

The commodification of health care provision, combined with the conception of clinician-agency and patient-principal, inculcates the notion of patient as consumer or client. This presents potentially two problems: First, it is conceptually suggestive that the ‘demand’ for health care (and health) is broadly comparable with the demand for commodities generally (as Le Grand 2003, cited above appears to suggest), and instils some echo with consumer sovereignty. This, as Galbraith (1973) notably argued, conflates the meaning of wants and needs: health care is frequently a fundamental need (although some mainstream health economists do explicitly recognise this distinction, they still tend to maintain the notion of health as a commodity, see Rice 1998). Wants, on the other hand, are not fundamental, and as Galbraith *et al.* have noted, may be manipulated by corporations and other parties in a materialistic society. Indeed, following this line of argument, pharmaceutical companies have an interest in creating new markets to perpetuate and expand demand. For instance, Reinhardt (2004) reports that pharmaceutical corporations’ outlays on marketing activities are approximately double research and development expenditure (see also Keaney 2002). Moreover, pharmaceutical and other medical supplies companies may be unintentional beneficiaries

of the marketing activities of corporations in the alcohol, food, and tobacco industries, which aims to increase consumption in those industries, but has a deleterious impact on health (Fine 2002; WHO 2002).

The wants-needs conflation is compounded by a further conflation within the information-theoretic agency approach. Theoretically, by generating greater information about the “value” of services and procedures agency advantage is eroded and the consumer is again sovereign. Such a notion is predicated on deeply flawed bases that there is no distinction between know-how and know-that, and needs are identical to wants and will be influenced by price. As noted, the experience of the Chinese reforms appears to demonstrate (Liu and Mills 2002) that the imposition of user charges may promote the notion of consumer sovereignty ahead of universality in health care provision (see also Iriart *et al* 2001; Price *et al* 1999).

Second, the separation between patients and clinicians implied by the commodification of health care and the resulting agency relationship may inadvertently undermine any promotion of preventive care activities. The potential patient as consumer may well be motivated differently from the potential patient as a member of a mutualised body. Simply, the former has an emphasis on individual consumption and exclusiveness, as opposed to universality (see also Keaney 2002).

The increasing prominence of efficiency concerns within the process of market-orientation accents the significance of measurable outcomes. Again this is a reflection of underlying theoretical framing based on instrumental rationality, where actions are devoid of values in themselves, only deriving value from their consequences, as in utilitarian philosophy.

The increased concentration on performance outcomes is engendering a

narrowing of the focus of incentives centring on finance (Fitzgerald 2004, Saltman 2002 *et al*). Liu and Mills (2002) have noted the monetisation of incentive structures in the Chinese system, and Hausman and Le Grand (1999) have expressed concerns that recent reform of the distribution of funds between general practitioner practices within primary care groups in the UK, offers potential for adverse incentive effects in the form of free-riding on the relative performance of other practices, which combined with “heavy monitoring” may engender more selfish behavioural traits. Motivation through monetisation associated with market-oriented reforms may be counter-productive in terms of eliminating corrupt activities.

Accompanying the contractualisation process is the importation of commercial accounting practices and mainstream health economic evaluation techniques, such as cost benefit analysis, into health care systems (Hildred and Watkins 1996, Keaney 2002, Saltman 2002, Gilson 2003 *et al*). This may be viewed as a useful means of generating more quantitative information, which will not only increase agents’ accountability, but also act as an objective means of evaluating those procedures producing the greatest net benefit, and hence rationing health services in an “efficient” fashion. This position presumes a relatively straightforward subjective-objective dual, and promotes a particular form of information as more ‘scientific’ without sufficiently recognising the framing effects involved in measuring, interpreting and judging data (see Hildred and Watkins 1996). To be sure, quantification relies on certainty, and confidence in the processes of measurement, yet many medical activities are profoundly uncertain and heterogeneous. An emphasis on measurable outcomes has the potential to distort activities in unintentional ways, and may relegate the process of health care (Daniels 1998).

A further manifestation of the contractualisation process that does not complement the quantitative emphasis of recent health policy is the impact this may have on trust (Gilson 2003; McMaster 2001). Recognition of the embeddedness of individuals in social systems analytically reveals the importance of trust between individuals, and individuals and institutions (Gilson 2003). Processes of contractualisation do not enhance the levels of trust between parties, tending to formalise relationships into specified channels. An outstanding issue for health policymakers is whether the pattern of health reform encouraged by supra-national bodies will increase the trustworthiness of health care institutions among the most vulnerable populations of the world, perhaps who have little trust in their states. The absence of trust between patients and the institutions of health care provision can, albeit unintentionally, provoke the agency problems alluded to earlier. An example perhaps of what Fine (2001) terms the iatrogenic consequences of (health) policy.

Commodification, the accompanying quantification of information, potential monetisation of incentives and possible corrosion of trust may also coalesce to change the character of medical care. Fitzgerald’s (2004) insightful anthropological study of different health care staff subject to New Zealand’s health care reforms provides food-for-thought. Fitzgerald found that concepts of care varied markedly between clinicians and managers: with the former tending to focus on the *person*, although consultants had the potential to view the patient as a scientific object. This had the potential to invoke feelings of empathy. By contrast managers tended to view care in the abstract, and as a homogenous entity that should be delivered routinely. Fitzgerald’s study suggests that as clinical workers experience time shortages, as a result of the rationalisation of medical

procedures, they experience disorientation and demotivation. They no longer feel they can fulfil the ethical requirements of their jobs, and conflict with managers is frequently observed (Fitzgerald 2002). In other words, the care process, as perceived by care workers, is crowded out by a different and potentially conflicting abstract view of care: there is a change in the values of the institution masquerading as “objective” science.

The recent global pattern of health policy has been informed by a laudable desire to address the inequities of health, while stressing the efficient provision of health care. It is the latter metric that is promoted by the complementary market-oriented reforms evident on a global scale. Yet the underlying basis for market-orientation is the subject of some concerns. The commodification of health care may undermine attempts to establish universality of provision, and focus on questionable measures of outcomes at the expense of the process of medical care. Indeed, in the latest survey of US health care markets Nichols, *et al* (2004) found “deep scepticism” about the ability of market-based reforms to improve the provision of medical care. Further research is required into ascertaining a more pluralistic basis for health policy, especially in the economics of health.

Selected References

- Baumol, William, G. (1993) “Health Care, Education and the Cost Disease”, *Public Choice*, 77, 17-28.
- Brinkerhoff, Derick. (2003) *Accountability and Health Systems: Overview, Framework and Strategies*. Bethesda, Maryland: Partners for Health Reformplus, Abt. Associates Inc.
- Daniels, Norman. (1998) “Symposium on the Rationing of Health Care 2: Rationing Medical Care—A Philosopher’s Perspective on Outcomes and Process”, *Economics and Philosophy*, 14, 27-50.
- Davis, John B. (2000) “Conceptualising the Lack of Health Insurance Coverage”, *Health Care Analysis*, 8, 55-64.
- Davis, John B. (2001) (Editor) *The Social Economics of Health Care*. London and New York: Routledge.
- Deaton, Angus and Darren Lubotsky. (2003) “Mortality, Inequality and Race in American Cities and States”, *Social Science and Medicine*, 56, 1139-1153.
- Department of Health. (2000) *The NHS Plan: A Plan for Investment, A Plan for Reform*, Cm4818-I, London: Stationery Office.
- DiMasi, Joseph A.; Ronald W. Hansen and Henry G. Grabowski. (2003) “The Price of Innovation: New Estimates of Drug Development Costs”, *Journal of Health Economics*, 22, 151-185.
- Diamond, Peter. (1998) “Symposium on the Rationing of Health Care 1: Rationing Medical Care—An Economist’s Perspective”, *Economics and Philosophy*, 14, 1-26.
- Enthoven, Alain C. (1994) “On the Ideal Market Structure for Third-Party Purchasing of Health Care”, *Social Science and Medicine*, 39, 1413-1424.
- Flood, Colleen M. (2000) *International Health Care Reform: A Legal, Economic and Political Analysis*. London and New York: Routledge.
- Fine, Ben. (2001) *Social Capital versus Social Theory: Political Economy and Social Science at the Turn of the Millennium*. London: Routledge.
- Fine, Ben. (2002) *The World of Consumption: The Material and Cultural Revisited* Second Edition. London: Routledge.
- Fitzgerald, Ruth. (2004) “The New Zealand Health Reforms: Dividing the Labour of Care”, *Social Science and Medicine*, 58, 331-341.

- Forget, Evelyn L. (2002) "National Identity and the Challenge of Health Reform in Canada", *Review of Social Economy*, 60, 359-375.
- Fuchs, Victor R. (1996) "Economics, Values and Health Care Reform", *American Economic Review*, 86, 1-23.
- Galbraith, John K. (1973) *Economics and the Public Purpose*. Harmondsworth: Penguin.
- Gilson, Lucy. (2003) "Trust and the Development of Health Care as a Social Institution", *Social Science and Medicine*, 56, 1453-1468.
- Grit, Kor and Wilfred Dolfsma. (2002) "The Dynamics of the Dutch Health Care System—A Discourse Analysis", *Review of Social Economy*. 60, 377-401.
- Hancock, Trevor. (1999) "Health Care Reform and Reform for Health: Creating a System for Communities in the 21st Century", *Futures*, 31, 417-436.
- Harding, April and Alexander S. Preker. (2003) (Editors) *Private Participation in Health Services*. Washington D. C.: World Bank.
- Hausman, Dan and Julian Le Grand. (1999) "Incentives and Health Policy: Primary and Secondary Care in the British National Health Service", *Social Science and Medicine*, 49, 1299-1307.
- Hildred, William and Larry Watkins. (1996) "The Nearly Good, the Bad, and the Ugly in Cost-Effectiveness Analysis of Health Care", *Journal of Economic Issues*, 30, 755-775.
- Hodgson, Geoffrey M. (1999) *Economics and Utopia: Why the Learning Economy is Not the End of History*. London & New York: Routledge.
- Inman, Robert P. (1987) "Markets, Governments, and the "New" Political Economy", in A.J. Auerbach and M. Feldstein (Editors), *Handbook of Public Economics*, Volume 2. Amsterdam: Elsevier, 647-777.
- Iriart, Celia; Amerson Merhy and Howard Waitzkin. (2001) "Managed Care in Latin America: The New Common Sense in health Policy Reform", *Social Science and Medicine*, 52, 1243-1253.
- Keaney, Michael (2002) "Unhealthy Accumulation: The Globalization of Health Care Privatization", *Review of Social Economy*, 60, 331-357.
- Khaleghian, Peyvand and Monica Das Gupta. (2004) *Public Management and the Essential Public Health Functions*. Washington DC: World Bank Policy Research Working Paper, 3220, 26pp.
- Le Grand, Julian. (2003) "Foreword", in April Harding and Alexander S. Preker, (Editors), *Private Participation in Health Services*. Washington D.C.: World Bank, pp ix-xi.
- Light, Donald W. (2001) "Managed Competition, Governmentality and Institutional Response in the United Kingdom", *Social Science and Medicine*, 52, 1167-1181.
- Liu, Xingzhu and Anne Mills. (2002) "Financing Reforms of Public Health Services in China: Lessons for Other Nations", *Social Science and Medicine*, 54, 1691-1698.
- Lloyd-Sherlock, Peter. (2000) (Editor) *Health Care Reform and Poverty in Latin America*. London: Institute of Latin American Studies.
- McMaster, Robert. (2001) "The National Health Service, the 'Internal Market' and Trust", in John B. Davis (Editor), *The Social Economics of Health Care*, London and New York: Routledge, 113-140.
- McMaster, Robert (2002) "A Socio-Institutionalist Critique of the 1990s' Reforms of the United Kingdom's

- National Health Service”, *Review of Social Economy*, 60, 403-433.
- Mooney, Gavin. (2001) “Communitarianism and Health Economics”, in John B. Davis (Editor), *The Social Economics of Health Care*, London & New York: Routledge, 40-60.
- Newhouse, Joseph P. (1993) “An Iconoclastic View of Health Cost Containment”, *Health Affairs*, (Supplement), 10, 152-171.
- Nicols, Len M.; Paul B. Ginsburg; Robert A. Berenson; Jon Christianson and Robert E. Hurley. (2004) “Are Market Forces Strong Enough to Deliver Efficiency Health Care Systems? Confidence is Waning”, *Health Affairs*, 23, 8-21.
- Niskanen, William A. (1968) “Nonmarket Decision Making: The Peculiar Economics of Bureaucracy”, *American Economic Review: Papers and Proceedings*, 58, 293-305.
- Pauly, Mark V. (1986) “Taxation, Health Insurance, and Market Failure in the Medical Economy”, *Journal of Economic Literature*, 24, pp.
- Price, David; Allyson M. Pollock and Jean Shaoul. (1999) “How the World Trade Organisation is Shaping Domestic Policies in Health Care”, *The Lancet*, 354, 1889-1892.
- Reich, Michael R. (2002) “Reshaping the State From Above, From Within, From Below: Implications for Public Health”, *Social Science and Medicine*, 54, 1669-1675.
- Reinhardt, Uwe. (2004) “An Information Infrastructure for the Pharmaceutical Market”, *Health Affairs*, 23, 107-112.
- Rice, Thomas. (1998) “The Desirability of Market-Based Health Reforms: A Reconsideration of Economic Theory”, in Morris L. Barer; Thomas E. Getzen and Greg L. Stoddart (Editors), *Health, Health Care and Health Economics: Perspectives on Distribution*. Hoboken: John Wiley & Sons.
- Sachs, Jeffrey, et al. (2001) *Macroeconomics and Health: Investing in Health for Economic Development*. Geneva: World Health Organization.
- Saltman, Richard B. (2002) “Regulating Incentives: The Past and the Present Role of the State in Health Care Systems”, *Social Science and Medicine*, 54, 1677-1684.
- Schut, Frederik T.; Stefan Greß and Juergen Wasem. (2003) “Consumer Price Sensitivity and Social Health Insurer Choice in Germany and the Netherlands”, *International Journal of Health Care Finance and Economics*, 3, 117-138.
- Sen, Amartya. (2002) “Why Health Equity?” *Health Economics*, 11, 659-666
- Smith, Peter C. (2002) “Performance Management in British Health Care: Will it Deliver?” *Health Affairs*, 21, 103-115.
- Stiglitz, Joseph. (1998) *More Instruments and Broader Goals: Moving Towards the Post Washington Consensus*. Helsinki: WIDER Annual Lecture.
- Sullivan, Terrence and Cameron Mustard. (2001) “Canada: More State, More Market?”, in John B. Davis (Editor), *The Social Economics of Health Care*, London & New York: Routledge, 172-192.
- Williamson, Oliver E. (2000) “The New Institutional Economics: Taking Stock, Looking Ahead”, *Journal of Economic Literature*, 38, 595-613.
- WHO. (World Health Organization) (2002) *World Health Report*. Geneva: World Health Organization.
- Websites**
- Organisation for Economic Cooperation and Development (OECD). www.oecd.org
- World Bank. worldbank.org
- World Health Organization. who.int

World Trade Organization. www.wto.org

*Robert McMaster
Department of Management
University of Glasgow, Scotland, UK
r.mcmaster@lbss.gla.ac.uk*

Health and Socioeconomic Status

Peter Muennig

Introduction

Socio-economic status is generally measured by an individual's income (or wealth), level of educational attainment, or occupation relative to others within the same country or social context. Each of these measures of socio-economic status is strongly associated with health indicators, such as life expectancy and infant mortality. For instance, in the United States, income-associated mortality appears to be the leading risk factor for death and disease, with over 360,000 deaths, 11 million years of life lost, and 17.4 million years of perfect health lost annually. In the United States, the very poorest have a life expectancy that is about 25% lower than the wealthiest, and the top 80% of income earners live over 4 years longer than everyone else (Rogot 1992; Muennig 2005). Similarly strong relationships have been found for education and occupation, and these effects have been noted across nations (Sorlie, Backlund et al. 1995; Marmot 2004).

The effects of socio-economic status on health are broader than the effects of abject poverty on health. In Europe, Japan, and (to a lesser extent) the US, relatively few citizens lack access to food, shelter, sanitation, clean water, and clothing. In many other nations, a significant portion of the population lacks one or more of these elements. This exposes such persons to an increased risk of death from deprivation alone. Because a large absolute number of persons are exposed to these conditions, abject poverty is likely the leading cause of death worldwide.

Relevant international public policy debates pivot around the role of development in alleviating the health outcomes of low socioeconomic status. Nonetheless, material resources cannot and should not be used to

define an individual's socio-economic status nor label that individual as poor. For instance, a Mexican landowner working in agriculture may be material resource poor relative to a federal employee at an airport (Black 2002). However, despite his lower earnings in agriculture, the landowner may well feel happier and more powerful working in the field than he would carrying tourists' luggage.

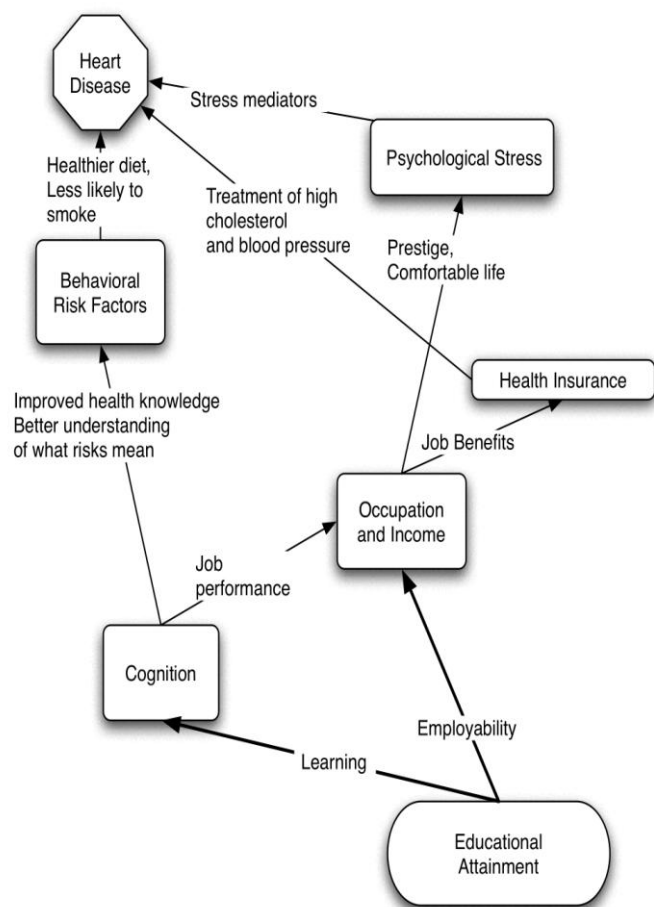
The study of the impact of socio-economic status on health focuses mostly on those with access to life's basic necessities, and mostly on industrialized nations. The focus on abject poverty is generally left to the field of international development. One side effect of this schism is that international development too often focuses on a person living on less than one dollar a day. Thus, mega projects (such as airports and dams), trade legislation, and other policies that improve earnings and open the doors to quick increases in gross domestic product are often emphasized over schools and public health. The potential side effects of such policies are an increase in perceived poverty, disempowerment, and a breakdown in cultural norms and civil society (Persell 1997). Even the psychological effects of these socio-cultural side effects can have a disruptive effect on health.

To understand how socio-economic status interacts with health, consider the way that risk factors for disease interact. At the most fundamental level, high blood pressure stresses the circulatory system leading to heart disease. This is referred to as a proximal (Latin for "close") cause of disease. But this high blood pressure arises in part from a mixture of genetic factors, diet, exercise, and psychological factors, such as stress (McEwen 1998). These factors in turn may partially relate to the individual's social support network, religion, and life stressors (such as housing quality, crime, and the work environment). These factors are in turn

influenced by fundamental risk factors for disease, such as laws assisting the poor and middle class.

Figure 1 provides a streamlined example of some of the ways that low educational attainment can lead to heart disease.

Figure 1. Links between Educational Attainment and Cardiovascular Disease.



In this diagram, we see that improving educational attainment improves cognitive decision-making. This in turn improves job performance. It also helps people better understand the risk of behavioral risk factors, such as smoking, drinking, and eating poorly. Likewise, a high school diploma or college degree improves one's earnings potential and increases the likelihood that their job will come with health insurance benefits (mostly only relevant in the United States and industrializing nations such as China). These factors in turn work to reduce psychological

stress, rates of smoking, improve dietary risk factors, and increase the likelihood that the person with higher educational attainment will receive anti-hypertensive medications or cholesterol lowering medications. In this figure, many links have been omitted for simplicity. For instance, reducing psychological stress may reduce smoking or eating poorly; many graduate students are well aware of the adverse effects of stress on these risk factors.

How Socioeconomic Status Affects Health

One way of thinking of high social class is that it provides information, power, and wealth that others cannot access and that the cumulative effect of this is better health (Link & Phalen 2005). From this perspective, all risk factors by socio-economic status are lumped together. For instance, persons with low educational attainment are also more likely to work in less prestigious jobs and have lower earnings than those who have higher attainment (Daly, Duncan et al 2002). They lack well-placed nepotistic connections or other resources needed to move up the social hierarchy. Such persons are therefore also more likely to live in high crime neighborhoods, occupy overcrowded and dangerous dwellings, have lower access to health care, eat less healthy foods, and engage in riskier health behaviors (Lantz et al 1998; Adler & Ostrove 1999; Kawachi et al 1999).

Link and Phelan's "fundamental cause" theory further argues that as risk factors for disease change, it will always be the wealthier members of society that survive because they have access to the basic social goods needed for survival (Link & Phalen 2005). Historically, risk factors for disease have shifted greatly within industrialized nations. Previously, infectious disease was the greatest risk factor, but this has been supplanted by chronic disease. Nonetheless, the poor have always been disproportionately affected.

Thus, the theory holds, low socio-economic status always has and always will be one of the strongest determinants of poor health and mortality.

Another way of looking at the relationship between socio-economic status and health, though, is to focus on the specific aspects of the health risks faced by persons of low socio-economic status in today's health climate. Income, education, and occupational class are each independent predictors of morbidity and mortality (Winkleby et al 1992). For instance, education may empower people to make better health choices regardless of their income (Fuchs 2004; Mechanic, 2002; Grossman 1997). Thus poorly educated people who have achieved relative financial success appear to be at higher risk of all cause mortality than others in their income category. Likewise, persons who have high educational attainment but lower earnings, such as university professors, may not be able to afford safer but more expensive cars, thereby increasing their risk of fatalities in motor vehicle accidents.

Despite some relatively clear links between socio-economic status and mortality, there is a good deal of controversy over what accounts for the majority of the 25% difference in life expectancy between the most fortunate and least fortunate members of society (Rogot 1992). If material differences, such as car quality and housing stock, were driving the association between socio-economic status and mortality, injuries should be dramatically higher among persons of low socio-economic status relative to those of high socio-economic status. Injuries do indeed drive differences in mortality by socio-economic status among children (Chen et al 2002). However, injuries due to crime or poor housing stock account for just 5% of deaths that create the yawning gap between socio-economic strata; most are due to cardiovascular disease (e.g., heart attacks),

cancer, diabetes, and infectious disease (Wong et al 2002).

The finding that these diseases are driving mortality differences by socio-economic status has lead to a search for differences in risk factors for cardiovascular disease among social classes. The most obvious candidates are a group of risks lumped under the term *behavioral risk factors*. These include smoking, poor eating habits, and excessive drinking. Other factors explaining this difference in heart disease rates might include *psychological risk factors*, *access to quality health care*, and *genetic risk factors*. One topic that has not received a good deal of attention is the contribution of *toxic or infectious exposures* to differences in heart disease by socio-economic status.

Behavioral Risk Factors

Behavioral risk factors were commonly believed to explain much of the difference in health status between the wealthy and the poor. There are good reasons for believing this. The prevalence of smoking among those with a bachelor's degree or more in 2000 was 11% relative to 32% among those with less than a high school diploma (National Center for Health Statistics 2002). Smoking is also an attractive explanation for lower mortality rates among Latinos, who have a smoking prevalence of 13% relative to 23% among non-Hispanic whites.

However, the hypothesis that behavioral risk factors explain the gap turns out to be a relative non-starter. Behavioral risk factors explain just 12-30% of the difference in mortality (Smith et al 1990; Lantz et al 1998). Given that these studies cannot capture the full effect of all behavioral risk factors combined (e.g., lower rates of seatbelt use or failure to invest in smoke alarms), behavioral risk factors may collectively play a larger role than is conventionally believed.

Income Inequality and Psychological Factors

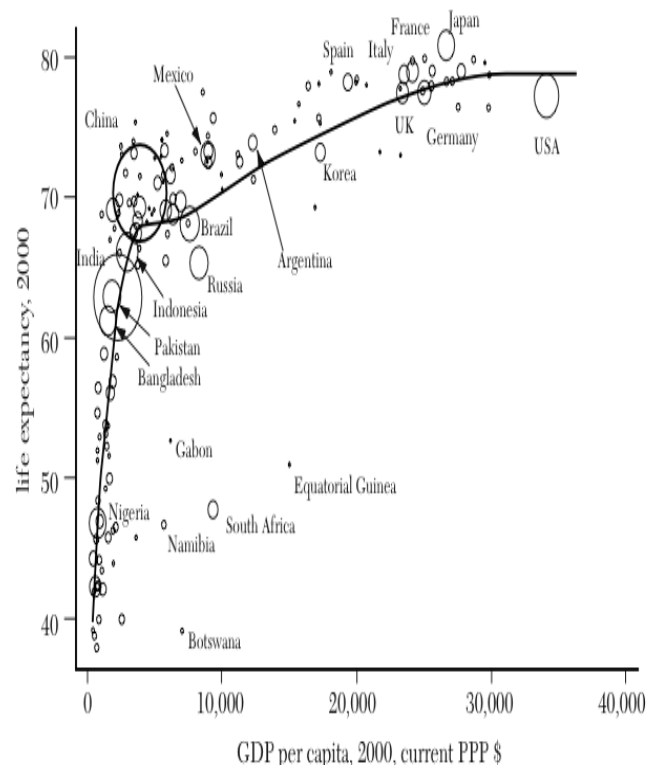
Income inequality refers to the gap between the wealthiest and poorest members of society, and is commonly measured using the Gini coefficient. A Gini of zero indicates that everyone shares the wealth equally, and a Gini coefficient of one indicates that one person holds all of the wealth. The Gini coefficient is a measure of *relative inequality*.

Another way of measuring income inequality is to estimate the resources needed for sustenance in a given nation and then estimate the number of people who live in households with incomes below this line (*absolute inequality*). This line is typically referred to as a *poverty line*. Because it is possible to have perfect equality, yet have 100% of the population below the poverty line, absolute inequality is perhaps better referred to as *absolute poverty*. Absolute poverty is a strong and logical predictor of mortality; the more people who cannot afford food and shelter, the higher the rate of mortality.

Relative income inequality has been proposed as a cause of mortality, possibly working by causing stress among those who are unfairly disadvantaged in the labor and education marketplace (Wilkinson 1999). Relative inequality garnered interest when the Whitehall study in the UK found that persons with high incomes near the top of the civil service ladder were at greater risk of mortality than those at the very top. Thus, even those with access to most of the material needs society can offer were inexplicably at greater risk of mortality than those in the highest positions (Smith et al 1990). This study has been replicated in many different countries, and the relationship is consistent across all measures of socio-economic status (e.g., persons with a Ph.D. are at lower risk of premature death than college graduates) (Sorlie et al 1995).

Preston (1976) was the first to note that the per capita wealth of a nation fails to predict mortality after a certain point. The health differences between the wealthiest and poorest members of society appear to be smaller in countries with more universal, rather than means-tested, social welfare policies, such as Sweden and Japan (Preston 1976). In fact, poor countries with large social investments (such as Costa Rica and Cuba) rank among the wealthiest in terms of life expectancy. Figure 2 is a classical Preston curve, shown below:

Figure 2. Classical Preston Curve.



Source: Adapted from World Bank (2002)

In this graph, life expectancy for the year 2000 is plotted against per capita gross domestic product (GDP). Those countries with a higher per capita GDP tend to have higher health until GDP reaches around \$3000 per person. After this point, the relationship flattens out somewhat. This suggests that factors other than economic prosperity become stronger predictors of

health once basic needs are met. Larger circles represent larger population sizes.

This figure outlines the relationship between per capita gross domestic product and life expectancy. Also intriguing is the finding that psychological and social milestones, such as a diploma, seem to be much more important than the actual number of years of schooling completed, even after controlling for the higher earnings conferred by such degrees (Backlund et al. 1999).

It has subsequently been noted that relative standing in society among both human (perceived perception) and non-human primates (actual social position) results in higher mortality as well as higher levels of stress hormone production and susceptibility to the common cold (Cohen et al 1997; Adler et al 2000; Goodman et al 2005; Sapolsky 2005). Over time, these higher levels of stress hormones may damage organ systems and bodily regulatory mechanisms, leading to higher rates of heart disease, diabetes mellitus, and infectious disease (McEwen 1998). These are precisely the conditions for which lower income persons are most at risk. Indeed, high levels of stress may even affect the genetic machinery that keeps cells young (Epel et al 2004).

Whether relative income inequality is associated with health nevertheless remains controversial for a number of reasons. First, income inequality in humans is ultimately a perception. Any relative comparison is, by definition, referenced to an individual's social and work environment. A wealthy stock trader in a corporation with a strong hierarchy may see herself as lower on the ladder than a highly regarded administrative assistant in a government job.

Moreover, while nearly anyone in low social standing will confirm that their position is stressful, it is difficult to separate the stress associated with perceived inequality from stress associated with other factors in the

lives of lower income persons (Taylor 2002). Indeed, as income or educational attainment increase, people are exposed to less crime, gain more control at work, spend less time in the car, and so forth. Thus, life's stressors decrease alongside any stress associated with social standing.

Finally, income inequality should have some relationship to differences in social standing within a society. However, data supporting or refuting the independent effects of income inequality are weak or non-existent (Lynch, Davey Smith et al. 2004). Moreover, the qualitative observation that countries with "egalitarian values" have higher life expectancy may have more to do with social programs or social values than with low rates of income inequality *per se*.

A broader definition of stress associated with low socio-economic status has been adopted that includes low control in the workplace, neighborhood, and home life (Marmot 2004). Other sources of stress include weaker social support networks and a less optimistic outlook on life (Ross and Wu 1996). Persons of lower socio-economic status are likely to report higher levels of stress due to all of these factors (Taylor 2002). Each of these sources of stress have been linked to cardiovascular disease, cancer, diabetes, and infectious disease (Lynch et al 1998; McEwen 1998; Everson et al 2001; Steptoe et al 2003; Steptoe et al 2003; Yan et al 2003). These sources of stress that may be unrelated to social position are relevant because low perceived social status may just be a marker for other psychological factors, such as insecurity, depression, or a lack of optimism.

Thus, income inequality *per se* may mostly be working as one of many psychological factors responsible for the health gradient. The implications for governance and policy that arise from the relative income hypothesis, some argue, are peculiar

(Mechanic 2002). For example, in the United States, medical innovations tend to benefit those with access to health care more than those that do not. Therefore, it could be argued that medical innovations are harmful to the health of the public because they exacerbate existing inequalities. The governance implications of absolute poverty (absolute inequality) are more concrete; redistribution of income through taxation will only produce improvements in health if the money is spent on programs—such as education, sanitation, or health care—that are made available to the poor.

Access to Quality Medical Care

In industrialized nations, access to medical care increases access to medications and treatments that are known to reduce morbidity and mortality. Of those diseases prominent in the health gradient—cardiovascular disease, cancer, diabetes, and infectious disease—access to medications that reduce cholesterol, blood pressure, and diabetes may be most important (Gregg et al. 2005). Moreover, it is early access to treatment that is perhaps most important for preventing mortality due to infectious diseases, such as pneumonia (Institute of Medicine, 2002).

However, cross-national studies show that universal access to healthcare plays a small role in predicting life expectancy (Reviewed in Lynch et al, 2004). In fact, despite a large literature demonstrating that many individual medical treatments have positive health effects and many experts agree that it works, it has never been conclusively demonstrated that the net effect of health insurance is beneficial (Hadley 2003). Part of the reason for this is that it is unethical to randomly assign subjects to receive no health insurance and wait to see if they are at greater risk of mortality than those who are randomly assigned health insurance.

The one randomized controlled study of health insurance, the 1982 Rand Health Insurance Experiment, therefore assigned 3,958 subjects to either receive a premium health insurance policy versus a policy that required financial contributions on the part of patients before they could receive care (Brook et al 1983). Subjects were then randomly assigned to receive a deluxe comprehensive free health care plan or a plan that required contributions on the part of the patient.

These authors found that mandatory patient contributions reduced health care utilization. This lower utilization appeared to lead to lower visual acuity and a 10% increased risk of death among high-risk subjects with hypertension. No improvements were found in 8 other measures of health. In sum, possessing a premium health plan greatly increased the use of medical care but had relatively little effect on morbidity and mortality.

This study is somewhat dated, so the past 20 plus years of advances in medical technology were not accounted for. It also only accounts for the difference in care provided to those with full access to the medical system relative to those with partial access, and therefore really only measures the impact of older preventive services and early treatment (Institute of Medicine, 2002). Indeed the one indisputably effective preventive modality available at the time, antihypertensive therapy, accounted for all of the difference in mortality between the experimental and control group. Since then, improved diabetes and cholesterol treatments have been introduced.

More recent studies show a 25 to 70% increase in mortality for any given age among the uninsured (Franks et al 1993; Sorlie et al 1994; Muennig et al 2005). These studies followed subjects over time, but did not randomly assign them to receive health insurance or no insurance. Rather, they used

statistical methods to predict how much better off the uninsured subjects would have been had they been given health insurance. Using this method, health insurance was found to account for 9 months of life expectancy. To put this in perspective, the uninsured are in an income category that predictive of a life expectancy in the ballpark of 4-8 years shorter than persons with income and education characteristics of the insured in the United States (Muennig et al 2005).

Some researchers suggest that observed improvements in life expectancy among the insured might in fact be explained by the economic protection that health insurance affords rather than the benefits associated with receiving medical attention (Ross and Mirowsky 2000). It is also possible that some of the association between health insurance and reduced mortality can be explained by unforeseen factors that incidentally happen to be associated both with better health and with health insurance. (This phenomenon is known as a third variable effect or endogeneity.)

In sum, in industrialized nations, health care likely plays a small but significant and growing role in reducing health disparities by socio-economic status (Muennig et al. 2005).

In developing nations, basic care primarily includes the treatment of infectious disease, which is a major cause of mortality among the poor. It is intuitive that small investments in medical care by governments produce larger gains in life expectancy, since such investments are typically limited to cost-effective modalities, e.g., vaccines, antibiotics, and dehydration solutions. Indeed, appropriate primary care is thought to be a critical component of basic public health infrastructure of developing nations (Mascie-Taylor and Karim 2003). However, even statements about the benefits of health care in the developing context must be highly qualified (Filmer & Pritchett 1999).

Genetic Factors

All individual characteristics either boil down to genetics or environmental influences. It follows, then, that if genetic factors were to blame for much of the differences in health by socio-economic status, then policy changes or educational interventions would have less effect on health outcomes. Such genes might impact educational attainment, occupation, earnings, and health simultaneously; an individual who does not have the biological capacity to succeed in school is unlikely to succeed in the business world or even in personal care, and there is little to be done about this (Gottfredson 2004).

An individual's genetic makeup can affect educational attainment by impacting personality characteristics (e.g., a lack of optimism), intellectual capacity (e.g., as measured by an intelligence quotient test), or other risk factors for disease. For instance, twin siblings reared apart show around a 50% concordance on measures of intelligence, and 30% of returns on education may be attributed to genetic factors (Miller et al. 1996; Lichtenstein & Pedersen 1997).

It is therefore possible that persons of lower IQ have worse health outcomes due to hampered health decision-making, or those with poor adaptive mechanisms have lower earnings and a higher biological susceptibility to stress. One possible pathway linking low innate intelligence to poor health outcomes is that a low innate intelligence sets off a cascade of social events such as the failure to develop peer networks, leads to a stressful work life, and so forth, which in turn reduces optimism and increases the risk of isolation and depression. Low innate intelligence may also reduce the uptake of preventive health messages and affect cognitive appraisal of dangerous situations. There is even circumstantial evidence to suggest that when the portion of the brain partially responsible

for memory is abnormally small, the stress response can be pathologically exaggerated (Gilbertson et al. 2002).

The ideal way to examine the effect of genetic factors on health and attainment is to randomly assign children to parents of varying socio-economic status. One study looked at circumstances in which this effectively occurs (Sacerdote 2004). In an innovative look at Korean adoptees who were essentially randomized to their adopting families, adoptees were just as likely as biological children to smoke or drink, and the chances of doing so were highly correlated with these risk factors in the adoptive parents. Biological and adopted children alike had a 19% increased risk of drinking and an 11% increase in smoking risk if their parents had these risk factors. Moreover, parental education was a strong predictor of educational attainment among adoptees. While children adopted by highly educated parents did not go on to do as well as their non-adopted siblings, children exposed to difficult circumstances generally do not fully catch up with their peers in either intelligence or health status after the difficulty is resolved et al. 1983; Chen et al. 2002).

However, it is highly unlikely that genetic factors have a large impact on either socio-economic status or health among poorer persons (Holtzman 2002; Sankar et al. 2004). Among persons of low socio-economic status, it is large social (environmental) obstacles that creates the slums and neighborhoods in which the majority of poor people live (Williams 1999; Fernandez and Su 2004). Individual genetics therefore do not likely exert much influence over the lower end of the health gradient simply because disabled and talented children alike are held back by environmental factors. Likewise, there is strong evidence that environmental stressors are much more prevalent among the poor, and show a gradient (Taylor 2002).

It is intuitive that, among students from wealthy households who are attending excellent schools, innate intelligence is a very strong determinant of academic success, a good career, and a healthy lifestyle. It is equally intuitive that innate intelligence plays little or no role in determining the future success and health status of children attending failing schools who have a single parent working two jobs. Indeed, this has been shown to be the case (Turkheimer et al. 2003). Other findings (e.g., that education is a stronger predictor of health status than IQ) also support this hypothesis (Link et al. 2003).

Other Factors

While cardiovascular disease and cancer explains most of the health gradient in industrialized nations, the risk for virtually every disease (exceptions include breast and uterine cancer), is substantially higher among persons of low socio-economic status (Lynch et al. 1996; Wong et al. 2002). The remainder of these factors can also be explained by stress, given indirect evidence that stress leads to premature cell aging (Epel et al 2004). In developing nations, the picture is somewhat murkier, with some countries showing a high risk of some more common diseases among the more affluent. Most of these diseases (e.g. heart disease and HIV) are related to the availability of cash to buy harmful goods and services, such as tobacco and the services of sex workers.

Nonetheless, the health gradient is seen in most studies conducted in developing nations as well (Liang et al. 2000). A large number of additional factors contribute to the health gradient in the developing context. A lack of sewerage or drinkable water, hunger, overcrowding, the lack of opportunities for exercise or health food consumption in poor neighborhoods, crime, unsafe work conditions, unsafe housing construction,

infectious diseases, and so forth appear to play a larger role in morbidity and mortality by socio-economic status.

Causal Evidence

For some risk factors associated with low socio-economic status, the direction of association is not clear. For instance, in the US, people may become ill because they cannot afford timely medical care, or people may not be able to afford timely medical care because they are ill, cannot work, and therefore cannot afford health insurance. Likewise, people may consume alcohol because they have an unpleasant, low-paying job or they may have a bad job because they drink.

One way to examine the direction of causality is via the use of prospective studies. In a prospective study, subjects who are healthy but poor are followed over time to see whether they are more likely to develop disease than those people who are healthy but wealthier.

Virtually all non-randomized prospective studies indicate that those who are healthy but poor are more likely to develop disease in old age than those who are healthy but rich. Such studies help establish that lower income mostly leads to disease rather than the other way around. For instance Lynch et al (1997) examined prospective data on healthy persons over a 25-year period. Relative to those without economic hardship measured over 3 time intervals, initially healthy subjects with sustained economic hardship have a 5.9 times higher risk of difficulties performing basic physical tasks later in life, while those with one or two episodes of hardship have a risk of 1.6 and 3.5 times higher, respectively. However, it is possible that some of these subjects were still poor over future intervals because they became ill and could not work. Had they not become ill, some subjects may

not have been poor at intervals 2 or 3 (Lynch et al. 1997).

Thus, one problem with non-randomized prospective analyses is that subjects have not been randomly assigned to their socio-economic status categories. It is therefore possible that something besides their socio-economic status is actually causing the poor health outcomes.

Indeed, because the incidence of chronic disease increases with age, older persons are especially at risk of economic hardship due to poor health. For instance, an examination of persons aged 50 and over in the US, the new onset of a chronic illness was found to reduce total wealth by 7% (Smith 1999). When reverse causality is inadequately controlled for in studies, the estimation of the total effect of socio-economic status on mortality will be overestimated. The impact of reverse causality is likely mitigated in the social welfare states of Europe and Asia because disability compensation rates tend to be higher; insurance against lost income due to disability will reduce the total effect of illness on wealth.

The gold standard for prospective studies is a randomized controlled trial. In this design, subjects are randomly assigned to receive benefits (such as income or housing) and are subsequently followed to see whether such benefits help or harm the individual. While some randomized controlled trials exist, prospective studies are in most circumstances the best available tool for studying relationships between socio-economic status and health (Kaplan & Keil 1993).

Other designs include natural experiments, such as before and after studies of groups that receive higher income or better access to educational opportunities. As discussed above, the Rand Health Insurance Experiment and the examination of Korean adoptees who were essentially randomly assigned to families generally meet these standards.

Additionally, the negative income tax experiments of the 1970s provide evidence that income supplementation is associated with improved birth outcomes. These studies examined the effect of randomly assigning low-wage income tax filers in various locations across the United States to receive money back on their tax returns versus just receiving standard welfare benefits as they had been before the study. The money back they received on low income tax filings was significantly larger than they would otherwise have received from standard welfare. One such study found a strong association between supplemental income and higher birth weight for high-risk pregnancies, with a gain of about 0.14 to 0.55 Kg (Kehrer & Wolin 1979).

Finally, an experiment conducted in 5 cities examined the impact of a randomized housing voucher experiment (Kling et al. 2004). Subjects receiving the vouchers tended to move to mixed income neighborhoods with improved living conditions. While no changes in general health status were noted over the 4 to 7 years of follow up, mental health and obesity rates fell among those receiving vouchers. The potential for long-term improvements in health is underscored by the fact that obesity and mental illness are both risk factors for heart disease.

One final type of study design is the natural experiment. Natural experiments entail studying the effects of a change in environmental conditions on health. Such studies are useful because they examine whether improvements in social conditions might reverse some of the negative effects of low socio-economic status on health.

Costello et al (2003), for example, examined the impact of redistribution from a Native American casino to the community. They found that social pathology—a major contributor to family ills among the poor in

the US—dropped significantly after income was redistributed (Costello et al. 2003).

In another careful study using synthetic cohorts, a strong correlation between the implementation of compulsory primary and secondary education in the United States and decreases in mortality was noted (Lleras-Muney 2004). Evidence for causality was also presented.

Not all such studies have positive findings. One study examined persons born in 1917 in the US who had their social security payments reduced relative to those born before 1917 due to a correction in the rate at which inflation was calculated. Those born in the last quarter of 1916 had 7-10% higher payments. When compared with the lower payment cohort, those with higher payments were found to have higher mortality. (Snyder & William 2002). The authors hypothesize that lower payments force elderly persons to work, thus reducing social isolation.

Contradicting these findings is a study using retirement survey data that shows each \$1000 annual increment in benefits lowers mortality by between 10-20 percent (Behrman et al. 1998). In a related South African study, large increases in social security payments made to elderly “black” and “colored” (ethnic South Asians) persons at the end of Apartheid produced significant gains in self-rated health (Case 2001). In this study, gains were realized across all members of the household because the money was shared with other family members.

Taking this evidence in its entirety, we see that the evidence for the association between socio-economic status and health is as strong as it is for smoking. In both instances, there is a large body of evidence showing an association, evidence of reversibility, multiple mechanisms, and limited alternative explanations (poor health reducing earnings in the case of income and risk taking personality traits among smokers). While the

magnitude of effects is likely large (especially for educational attainment where there is no strong alternative explanation), it is difficult to quantify precisely.

International Differences

As discussed above, there tends to be a strong relationship between per capita gross domestic product (GDP) and life expectancy up to approximately \$3000 US per year. (See Figure 2.) After this point, the relationship becomes much weaker as indicated by the flattening of the line. This suggests that purchases that benefit health (either by governments or individuals) are optimized once a nation reaches a given level of economic development. Looking again at this figure, we see that there are many poor nations with low per capita GDP but high life expectancy. It is therefore conceivable that some nations are making better health purchasing decisions.

These nations include Costa Rica, Cuba, and Chile. In addition, Kerala (a state in India) has unexpectedly high life expectancy, an observation that some feel is attributable to heavy investments in education, particularly of females (Sen 1993). Some researchers suggest that social welfare policies (such as universal healthcare, large educational investments, and/or income redistribution programs) explain why these poorer countries or states have achieved life expectancies comparable to those of much wealthier ones. Similarly, a wide range of European nations have higher life expectancies than the US despite having half to two-thirds the per capita GDP.

Nonetheless, differences in culture, climate, corruption, conflict, and other factors that can have an impact on health may explain the observed differences in life expectancy. Regression models generally find that education explains much of the variation in life expectancy between nations, but other

social welfare investments are important as well (Reviewed in Lynch et al, 2004). While subjective comparisons and statistical models do lend evidence to the hypothesis that social investments lead to differences in life expectancy, they cannot be considered conclusive.

In these studies, income inequality is not a predictor of mortality once other known social factors such as educational attainment are controlled for. It does, however, seem to be a predictor of crime rates and homicide (Hsieh and Pugh 1993). Small area studies within the US at one point in time show a positive but moderate to weak association between income inequality and “all cause” mortality.

While some see economic globalization as a positive force that has potential to reduce conflict between nations and increase prosperity, others point to perils that might harm the health of the world’s poor. These include shifts in manufacturing from countries with rigorous occupational safety and environmental standards to nations in which such standards are virtually nonexistent (Shaffer et al. 2005). Other issues, such as child labor, are also prominent in the public health debate. International institutions that regulate trade have the potential to influence developing nations’ domestic public health policies, and thus have a strong influence over the health and socio-economic status of their residents.

Differences Across Time

Some researchers have looked to historical data to ascertain whether specific social investments, economic depressions, or other events might lead to changes in mortality. Amartya Sen notes that the lowest civilian mortality rates in England and Wales occurred in the decades 1911-1921 and 1940-1951. Over these decades, the World Wars forced the government to massively increase

progressive taxation and redistribute food and medicine while creating jobs for the bulk of the population, especially females. While the growth in life expectancy from 1900-1960 otherwise ranged from 1.4% to 4%, the decade containing World War I saw a 6.5% improvement. This improvement was undoubtedly diminished by the great influenza pandemic of 1918-19. The decade encompassing World War II saw an increase of 6.8%. Increases also occurred among elderly persons following the institution of Social Security in the 1930s and Medicare in the 1960s. Many factors play a role in the change of disease incidence over time, limiting conclusions that can be drawn from such trends.

Governance Implications

If the lower civilian mortality rates in Britain that occurred during both world wars was actually due to redistributive effects, it would make sense for every world leader to maximize these policies. Qualitative cross-national analyses back this presupposition, suggesting that countries with large scale social programs do better with respect to life expectancy than those that fail to provide adequate benefits to lower income citizens. This is perhaps most dramatically illustrated by Cuba, Chile, Costa Rica, the Indian state of Kerala, and pre-conflict Sri Lanka, all of which produced a life expectancy similar to the United States despite a much lower per capita gross domestic product. The evidence cited in this entry supports these qualitative assessments.

In the UK, a report on the impact of health inequalities, known as the Acheson report, concluded that larger investments in social programs would improve the health of England (Acheson 2002). Presumably, these health gains would be accompanied by economic gains that would at least partly offset the cost of such programs. The many

potential benefits of such programs include healthier, more productive workers, a better-educated workforce, and reductions in morbidity and mortality.

One tenet of the Acheson report is that government agencies need to work together to provide more comprehensive, evidence-based social interventions. The idea is that the sum total effect of two services is greater than one. For instance, job training will be much more effective if delivered with daycare. It may be more effective still if bundled with counseling in interview skills and access to appropriate attire. Because agencies are required to collaborate, it might also be reasonable to expect that services would be delivered more efficiently.

Some of the recommendations of the Acheson report have subsequently been implemented (Marmot 2004). However, if one accepts the premise that social programs improve health, fixing social ills is no easy matter. As Mechanic (2002) notes, simply taking money from wealthy people and giving it to poor people is not likely a viable solution. The challenge, then, is to come to a reasonable consensus on which programs build (or at worst maintain) a sense of autonomy among the recipient while still repairing the underlying mechanism causing the poor health.

Another major challenge lies in human resource management and organizational change management for government programs. In many instances, even highly trained social workers may not fully believe in their potential to change lives, and may therefore see their job as unrewarding. Incentives for employee advancement and continuous quality improvement are too often overlooked in the public sector. These problems can sometimes be overcome with relatively simple and straightforward changes in the structure of work (Tendler 1998).

A final problem with large-scale social programs is whether the targeted recipients actually want or care about the services. A common complaint of inner city schoolteachers in the United States, even those in well-funded schools, is that the kids do not show up for class and the parents fail to meet up with the teachers. Part of the reason for this is clear; students don't enjoy school (especially when they aren't learning basic reading and writing skills) and parents are too busy working two jobs and dealing with a wide array of stressful social problems to attend to their children's educational needs. For some parents, education may not been seen as an important asset. Rather, it may be seen as just another obligation that the family has to meet for a society that has treated them unfairly (Lewis 1966). Recipients of government services need to see these services as a cherished commodity. One way of doing this is to design such programs with the participatory input of these programs' clients. Another way to accomplish this is to limit interventions geared to enhancing existing social institutions to those who show up first. Eventually, those who do not have access to such services may come to demand them.

However, many recipients of government services are already demanding that they be fixed, but they lack a voice in government. In rare cases, these groups do manage to get the attention of local politicians. For instance, a group of high school students in the South Bronx in New York City has formed an effective lobby group that is demanding that the failing schools they attend be repaired (Su 2005).

It is perhaps ironic that those demanding improvements in services are so rarely heard; one central idea of such government programs is that they offer a hand up to beneficiaries, providing them with the human capital needed to climb in social class

(Sorokin 1959). In the theoretical ideal, societies are constructed as "meritocracies", in which those from lower social strata have the opportunity to climb social ladders provided that they are willing to work for it. In practice, there are many psychological and social barriers to doing so, even in modern industrialized nations.

There is evidence that compulsory schooling laws greatly facilitated social mobility and may have had a large impact on human life expectancy in the United States and elsewhere (Lleras-Muney 2004). This has allowed a good number of individuals to break free from many generations of low social status. However, in most countries, compulsory education has failed to meet its potential as a tool for building human capital and social mobility because the schools in low-income communities are either of low quality or are non-existent.

By this logic, it follows that effective education interventions designed to improve the quality of schooling in low-income neighborhoods will likely improve future earnings, improve sense of self-efficacy, reduce crime, and improve social networks of low income beneficiaries (Karoly & Bigelow 2005; Reynolds 2001). Improved earnings and occupational status, coupled with improved decision-making, in turn translate into health benefits.

Improvements to the educational system in any country's low-income communities might include: infrastructure improvements, reduced class sizes, support- and oversight-intensive charter school expansion, pre-kindergarten programs, bilingual programs, apprenticeship programs, teacher salary enhancements, teacher-peer mentoring programs, standardized exam preparatory courses, school libraries and public library branches, and after-school tutoring programs. Such enhancements come at a considerable cost,

yet their efficiency is mostly measured using improvements in earnings alone.

While the improvements in earnings associated with schooling are large, they may not be large enough to justify very comprehensive (and expensive) social programs. A recent RAND study in the state of California in the United States examined some non-labor market benefits associated with pre-kindergarten programs, such as reduced crime. This study found returns in the range of \$2-\$4 per dollar invested (Karoly and Bigelow 2005). Thus, there is evidence that educational enhancements can produce health benefits alongside other social benefits while reducing long-term societal costs.

Some argue that these benefits can be realized without up front costs. For instance, improving human resource management of the teacher workforce could potentially accomplish the same goals as salary augmentation.

Because it is likely that educational enhancements come at low long-term costs or even savings, this route of social resource redistribution is perhaps most attractive on face value. However, other large scale, well-executed social programs such as health insurance may also produce net social benefits that go beyond health as well. For instance, an analysis by the non-partisan US Congressional Budget Office in the United States found that a single payer health insurance program would reduce the total health bill, thus potentially lowering worker contributions (Congressional Budget Office 1993).

As evidenced by experiments cited in *Community Health and Medicine*, successful community-based interventions tend to be participatory and address many different social problems at once. Some of these programs have radically transformed the communities within which they were implemented.

In 1962, Bolivar County was mostly made up of dormant farmland occupied by 14,000 people who were displaced by the mechanization and globalization of cotton crops. The community reported a median income of \$900 per year (about \$5,000 constant 2004 dollars), a median educational attainment of 5 years, and an infant mortality of 70 deaths/1000 live births (relative to 26 deaths/1000 births nationwide that year). The ambitious community revitalization project joined a multifaceted education intervention with housing improvements, the construction of a cooperative farm (which exported food and hired community members), legal services, financial services, and health care with extensive outreach and health education programs. The intervention was a phenomenal success, moving Bolivar County within the range of national means for income, educational attainment, and infant mortality for African Americans (Geiger 2002).

The Bolivar County intervention is especially notable given that it occurred in an industrialized nation where implementation costs were high, but nonetheless reduced infant mortality fivefold while radically changing the socio-demographic and economic landscape (including the labor market) of the community.

In the United Kingdom, the Sure Start program serves as a national version of the Bolivar County experiment. This is a £540 million program (US\$1.026 billion) targeted toward disadvantaged families with pre-school aged children (Roberts 2000). It includes outreach and support for families, pre-school education interventions, health care and child development advice, skills training, debt counseling, and literacy training.

Health outcomes for this program will likely be available in a few years. Perhaps more importantly, initiatives focus in part on

parental engagement as an outcome measure. Thus, this program measures whether it has internal validity (is getting parents excited about using its resources and is delivering content efficiently) as well as whether it achieves larger scale objectives such as improving health outcomes.

This and other multi-faceted interventions with an evaluation component are ultimately needed before large-scale programs designed to help people out of poverty can be widely implemented. Previous experiments, such as the negative income tax experiments of the 1970s in the United States, may have been hampered by failing to fully consider individual agency and autonomy (Kehrer & Wolin 1979).

Another approach to improving health outcomes among socio-economically disadvantaged populations is to bypass questions of autonomy altogether. For instance, Link and Phelan (2005) argue that the fundamental cause of socio-economic disparities in health is that low-income groups lack access to information, prestige, social connections, and power that optimize survival. Therefore, as new risks arise, low-income populations are the slowest to adapt. Given this, one policy approach is to simply bypass solutions that require individual action and instead focus on those for which no action is needed on the part of the individual (Link & Phalen 2005). For instance, rather than focusing on educating families about the dangers of lead paint, it is more effective to simply abate all lead paint through legislative action. This approach, however, fails to address the underlying cause of inaction: a lack of human capital among program recipients.

A final approach to improving the health of disadvantaged populations is based on intensively targeting a given risk factor within the low-income population. For instance, anti-smoking media campaigns have reduced

smoking rates among minority groups in the UK (Lowey et al. 2003).

In most nations, most people to the political right or left of the political spectrum genuinely wish to improve the health and quality of life of their fellow citizens. The disagreement centers on how to go about solving problems. Amalgamating both perspectives using evidence-based policymaking may be critical to accomplishing this shared goal.

There is evidence that large government programs can be quite expensive and potentially harmful if they are implemented incorrectly. Likewise, labor market forces are often not considered.

Similarly, much needs to be learned about delivering non-market services, such as health care and education. Despite advice to the contrary from introductory economics textbooks, too often governments try to privatize health care, education, and other forms of social service delivery.

Healthcare provides the most dramatic example of the failings of this approach. The United States has now built nearly 30 years of experience into attempts at making health care work within the private sector, and has only succeeded in delivering services to a subset of the population at over of twice the cost per capita of industrialized nations with universal health care (National Center for Health Statistics 2002).

A final challenge facing large-scale interventions designed to reduce the health and social impact of low socio-economic status populations is globalization. As low skilled jobs get outsourced from industrialized nations to poorer nations, the pool of unskilled jobs within these industrialized nations is shrinking. While globalization is improving information exchange, increasing global economic growth, and producing other benefits, those without advanced technical skills living in

industrialized nations may increasingly become socially isolated.

These trends simply underscore the need to maximize all citizens' chances of participating in the global economy by building effective governmental institutions, optimizing opportunities for economic participation and success, and improving neighborhood and social characteristics.

In turn, programs geared toward accomplishing these goals can benefit greatly by turning to the private sector for lessons in organizational change management and human resources management. Finally, macro-economic policies need to be more rationally guided. Governmental regulations can stifle economic growth and should be used prudently.

Examples of rational market regulations that improve the health of the socio-economically disadvantaged include child labor laws, environmental protections, and occupational safety laws. Irrational regulations include price fixing, some trade barriers, and gasoline subsidies, all of which can harm the poorest members of society by slowing economic growth and sapping resources from more efficient social programs. This points to one final solution: training social scientists in economics, and training economists in the social sciences.

Selected References

- Acheson, Donald. (2002) *Independent Inquiry into Inequalities in Health Report*. London: the Stationary Office: 164.
- Adler, Nancy E; E.S. Epel; and G. Castellazzo. (2000) "Relationship of Subjective and Objective Social Status With Psychological and Physiological Functioning: Preliminary Data in Healthy White Women", *Health Psychology*, 19, 6, 586-92.
- Adler, Nancy E. and Joan M. Ostrove. (1999) "Socioeconomic Status and Health: What We Know and What We Don't", *Annals of the New York Academy of Science* 896: 3-15.
- Backlund, Eric; P.D. Sorlie and N.J. Johnson. (1999) "A Comparison of the Relationships of Education and Income With Mortality: the National Longitudinal Mortality Study", *Social Science and Medicine*, 49, 1373-1384.
- Black, Maggie (2002) *The No-Nonsense Guide to International Development*. Scranton, PA: Verso.
- Brook, Robert H.; J.E. Ware Jr; W.H. Rogers; E.B. Keeler; A.R. Davies; C.A. Donald; G.A. Goldberg; K.N. Lohr; P.C. Masthay and J.P. Newhouse. (1983) "Does Free Care Improve Adults' Health? Results From a Randomized Controlled Trial", *New England Journal of Medicine*, 309, 23, 1426-34.
- Case, A. (2001) *Does Money Protect Health Status? Evidence From South African Pensions*. National Bureau of Economic Research. New York: NBER.
- Congressional Budget Office. (1993) *Single-Payer and All-Payer Health Insurance Systems Using Medicare's Payment Rates*. Washington, DC: Congressional Budget office.
- Chen, Edith; K.A. Matthews and W.T. Boyce. (2002) "Socioeconomic Differences in Children's Health: How and Why Do these Relationships Change With Age?" *Psychol Bull*, 128, 295-329.
- Cohen, Sheldon; Scott Line; Stephen B. Manuck; Bruce S. Rabin; Eugene R. Heise and Jay R. Kaplan. (1997) "Chronic Social Stress, Social Status, and Susceptibility to Upper Respiratory infections in Nonhuman Primates", *Psychosomatic Medicine*, 59, 3, 213-21.
- Costello, E. Jane; Scott N. Compton; Gordon Keeler and Adrian Angold. (2003) "Relationships Between Poverty and Psychopathology: A Natural Experiment", *JAMA*, 290, 15, 2023-9.

- Daly, Mary C. et al. (2002) "Optimal indicators of Socioeconomic Status for Health Research", *American Journal of Public Health*, 92, 7, 1151-7.
- Davey-Smith, George et al. (1990) "Magnitude and Causes of Socioeconomic Differentials in Mortality: Further Evidence From the Whitehall Study", *Journal of Epidemiology and Community Health*, 44, 4, 265-70.
- Epel, Elissa S. et al. (2004). "Accelerated Telomere Shortening in Response to Life Stress", *Proceedings of the National Academy of Science USA*, 101, 49, 17312-5.
- Everson, S. A. et al. (2001). "Stress-induced Blood Pressure Reactivity and incident Stroke in Middle-Aged Men", *Stroke*, 32, 6, 1263-70.
- Fernandez, Roberto and Celina Su (2004) "Space in the Study of Labor Markets", *Annual Review of Sociology*, 30, 545-569.
- Filmer, Deon and Lant Pritchett (1999). "the Impact of Public Spending on Health: Does Money Matter?", *Social Science and Medicine*, 49, 10, 1309-23.
- Franks, Peter et al. (1993). "Health insurance and Mortality: Evidence From A National Cohort", *JAMA*, 270, 6, 737-41.
- Fuchs, Victor (2004) "Reflections on the Socio-Economic Correlates of Health", *Journal of Health Economics*, 23, 653-661.
- Geiger, H. Jack (2002) "Community-Oriented Primary Care: A Path to Community Development", *American Journal of Public Health*, 92, 11, 1713-6.
- Gilbertson, Mark W. et al. (2002). "Smaller Hippocampal Volume Predicts Pathologic Vulnerability to Psychological Trauma", *Nature Neuroscience*, 5, 11, 1242-7.
- Goodman, Elizabeth et al. (2005). "Social Inequalities in Biomarkers of Cardiovascular Risk in Adolescence", *Psychosomatic Medicine*, 67, 1, 9-15.
- Gottfredson, Linda S. (2004). "intelligence: is it the Epidemiologists' Elusive 'Fundamental Cause' of Social Class inequalities in Health?", *Journal of Personality and Social Psychology*, 86, 1, 174-99.
- Gregg, Edward W. et al. (2005). "Secular Trends in Cardiovascular Disease Risk Factors According to Body Mass index in US Adults", *JAMA*, 293, 15, 1868-74.
- Hadley, Jack (2003). "Sicker and Poorer: the Consequences of Being Uninsured: A Review of the Research on the Relationship Between Health Insurance, Medical Care Use, Health, Work, and Income", *Medical Care Research and Review*, 60, 2, Supplement, 3S-75S; Discussion 76S-112S.
- Holtzman, Neil A. (2002). "Genetics and Social Class", *Journal of Epidemiology and Community Health*, 56, 7, 529-35.
- Hsieh, Ching-Chi and Mark D. Pugh (1993) "Poverty, income inequality, and Violent Crime: A Meta-Analysis of Recent Aggregate Data Studies", *Criminal Justice Review*, 18, 182-202.
- Institute of Medicine (2002) *Care Without Coverage: Too Little, Too Late*. Washington, DC: National Academies Press, Institute of Medicine, 2004.
- Jere C. Behrman et al. (1998) *Causes, Correlates, and Consequences of Death Among Older Adults: Some Methodological Approaches and Substantive Analyses*. Boston: Kluwer Academic Publishers.
- Kaplan, George A. and J. E. Keil (1993) "Socioeconomic Factors and Cardiovascular Disease: A Review of the Literature", *Circulation* 88(4 Pt 1): 1973-98.
- Karoly, Lynn A. and James E. Bigelow (2005). The Economics of investing in Universal Preschool Education in California. *Rand Labor and Population*. Santa Monica, Rand Corporation, 238.
- Kawachi, Ichiro et al. (1999) "Crime; Social Disorganization and Relative Deprivation", *Social Science and Medicine*, 48, 6, 719-31.

- Kehrer, Barbara H. and Charles M. Wolin (1979). "Impact of income Maintenance on Low Birth Weight: Evidence From the Gary Experiment", *Journal of Human Resources*, 14, 4, 434-62.
- Kling, Jeffrey et al. (2004) *Moving...to Opportunity... and... Tranquility: Neighborhood Effects on Adult Economic Self-Sufficiency and Health From A Randomized... Housing... Voucher Experiment*. New York: National Bureau of Economic Research, KSG Working Paper: 57.
- Lantz, Paula M. et al. (1998). "Socioeconomic Factors, Health Behaviors, and Mortality: Results From A Nationally Representative Prospective Study of US Adults", *JAMA*, 279, 21, 1703-8.
- Lewis, Oscar (1966). *La Vida: A Puerto Rican Family in the Culture of Poverty*. London: Panther.
- Liang, Jersey et al. (2000). "Socioeconomic Gradient in Old Age Mortality in Wuhan, China", *J. Gerontol B Psychol Sci Soc Sci*, 55, 4, S222-33.
- Lichtenstein, P. and N. L. Pedersen. (1997) "Does Genetic Variance For Cognitive Abilities Account For Genetic Variance in Educational... Achievement... and Occupational Status? A Study of Twins Reared Apart and Twins Reared together", *Social Biology*, 44, 1-2, 77-90.
- Link, Bruce G. and Jo. Phelan... (2005) in Mechanic, D. *Policy Challenges in Modern Health Care*. New Brunswick, N.J.: Rutgers University Press.
- Link, Bruce G. et al. (2003). *The Resources That Matter: Fundamental Social Causes of... Disease and the... Challenge of intelligence*. Paper Presented At the Meetings of the American Sociological Association, Atlanta, GA.
- Lleras-Muney, Adriana... (2004). *The Relationship Between Education and Adult Mortality in the United States*. PhD thesis. New York: Columbia University.
- Lowey, H., K. toque et al. (2003) "Smoking Cessation Services Are Reducing Inequalities", *Journal of Epidemiology and Community Health*, 57, 8, 579-80.
- Lynch, John et al. (2004) "Is Income Inequality A Determinant of Population Health? Part 1. A Systematic Review", *Milbank Quarterly*, 82: 5-99.
- Lynch, John et al. (1998) "Does Low Socioeconomic Status Potentiate the Effects of Heightened Cardiovascular Responses to Stress on the Progression of Carotid Atherosclerosis?" *American Journal of Public Health*, 88, 3, 389-94.
- Lynch, John et al. (1996). "Do Cardiovascular Risk Factors Explain the Relation Between Socioeconomic Status, Risk... of... All-Cause... Mortality, Cardiovascular... Mortality, and Acute Myocardial infarction?" *American Journal of Epidemiology*, 144, 10, 934-42.
- Lynch, John et al. (1997). "Cumulative Impact of Sustained Economic Hardship on Physical, Cognitive, Psychological, and Social Functioning", *New England Journal of Medicine*, 337, 26, 1889-95.
- Marmot, Michael G. (2004). "Tackling Health Inequalities Since the Acheson inquiry", *Journal of Epidemiology and Community Health*, 58, 4, 262-3.
- Mascie-Taylor, Nicholas C. G. and Enamul Karim. (2003) "The Burden of Chronic Disease", *Science*, 302, 5652, 1921-2.
- McEwen, Bruce S. (1998). "Protective and Damaging Effects of Stress Mediators", *New England Journal of Medicine*, 338, 3, 171-9.
- Mechanic, David (2002) "Disadvantage, inequality, and Social Policy", *Health Affairs (Millwood)*, 21, 2, 48-55.
- Miller, Paul et al. (1996). "Earnings and Schooling: An Overview of Economic Research Based on the Australian Twin Register", *Acta Geneticae Medicae Et Gemellologiae*, 45, 4, 417-29.
- Money, John et al. (1983) "Growth of intelligence: Failure and Catch-Up Associated Respectively With Abuse and Rescue in the Syndrome of Abuse

- Dwarfism", *Psychoneuroendocrinology* 8, 3, 309-19.
- Muennig, Peter et al. (2005) "The Cost Effectiveness of Health Insurance", *American Journal of Preventive Medicine* 28, 1, 59-64.
- National Center For Health Statistics, Health, United States, 2002. *With Chartbook on Trends in the Health of Americans*. Hyattsville, Maryland: 2002.
- Persell, Caroline. (1997). "The Interdependence of Social Justice and Civil Society", *Sociological Forum*, 12, 2, 149-172.
- Preston, Samuel H. (1976) *Mortality Patterns in National Populations: With Special Reference to Recorded Causes of Death*. New York: Academic Press.
- Roberts, Helen (2000) "What is Sure Start?" *Archives of Disease in Child* 82, 6, 435-7.
- Rogot, Eugene (1992) *A Mortality Study of 1.3 Million Persons By Demographic, Social, and Economic Factors: 1979-85 Follow-Up*. Bethesda, MD: National Institutes of Health.
- Ross, Catherine E. and John Mirowsky (2000). "Does Medical insurance Contribute to the Socio-Economic Differentials in Health?" *Milbank Quarterly*, 78, 291-321.
- Ross, Catherine E. and Chia-Ling Wu (1996) "Education, Age, and the Cumulative Advantage in Health", *Journal of Health and Social Behavior*, 37, 1, 104-20.
- Sacerdote, Bruce. (2004) *What Happens When We Randomly Assign Children to Families?* New York: National Bureau of Economic Research.
- Sankar, Pamela et al. (2004). "Genetic Research and Health Disparities", *JAMA* 291(24): 2985-9.
- Sapolsky, Robert M. (2005). "The Influence of Social Hierarchy on Primate Health", *Science* 308(5722): 648-52.
- Sen, Amartya (1993) "The Economics of Life and Death", *Scientific American* 268, 5, 40-7.
- Shaffer, Ellen et al. (2005). "Global Trade and Public Health", *American Journal of Public Health*, 95, 1, 23-34.
- Smith, James P. (1999) "Healthy Bodies and Thick Wallets: the Dual Relation Between Socio-Economic Status and Health", *J Economic Perspectives*, 13, 145-166.
- Snyder, Stephan and William Evans. (2002) *The Impact of Income on Mortality: Evidence From the Social Security Notch*. New York: National Bureau of Economic Research: 53.
- Sorlie, Paul D. et al. (1995). "US Mortality By Economic, Demographic, and Social Characteristics: the National Longitudinal Mortality Study", *American Journal of Public Health*, 85, 7, 949-56.
- Sorlie, Paul D. et al. (1994) "Mortality in the Uninsured Compared With That in Persons With Public and Private Health Insurance", *Archives of Internal Medicine*, 154, 21, 2409-16.
- Sorokin, Pitirim. (1959) *Social and Cultural Mobility*. New York: The Free Press.
- Steptoe, Andrew et al. (2003) "influence of Socioeconomic Status and Job Control on Plasma Fibrinogen Responses to Acute Mental Stress", *Psychosomatic Medicine* 65, 1, 137-44.
- Steptoe, Andrew, G. Willemsen et al. (2003) "Socioeconomic Status and Hemodynamic Recovery From Mental Stress", *Psychophysiology*, 40, 2, 184-91.
- Su, Celina. 2005. *Reading, Writing, and Reform in the South Bronx: Lessons For Family-School Partnerships*. [Research Digest]. Cambridge, MA: Harvard Graduate School of Education, Harvard Family Research Project.
- Taylor, Humphrey. (2002). *Poor People and African-Americans Suffer the Most Stress From the Hassles of Daily Living*. the Harris Poll, Harrisinteractive, 5.
- Tendler, Judith. (1998). *Good Government in the Tropics*. Baltimore: Johns Hopkins University Press.
- Turkheimer, Erik, A. et al. (2003) "Socioeconomic Status Modifies

Heritability of IQ in Young Children”,
Psychological Science, 14, 6, 623-628.....
 Wilkinson, Roy G. (1999). “Health,
 Hierarchy, and Social Anxiety”, *Annals of
 the New York Academy of Science*, 896:
 48-63.
 Williams, David R. (1999). “Race,
 Socioeconomic Status, and Health: the
 Added Effects of Racism and
 Discrimination”, *Annals of the New York
 Academy of Science*, 896, 173-88.
 Winkleby, Marilyn et al. (1992)
 “Socioeconomic Status and Health: How
 income, Occupation, and Education
 Contribute to Risk Factors for
 Cardiovascular Disease”, *American
 Journal of Public Health*, 82, 816-820.
 Wong, Mitchell et al. (2002) “Contribution of
 Major Diseases to Disparities in
 Mortality”, *New England Journal of
 Medicine*, 347, 20, 1585-92.
 Yan, Lijing et al. (2003). “Psychosocial
 Factors and Risk of Hypertension: The
 Coronary Artery Risk Development in
 Young Adults (CARDIA) Study”, *JAMA*,
 290, 16, 2138-48.

Webites

MacArthur Network on SES and Health.
www.macses.ucsf.edu
 Health, United States of America.
www.cdc.gov/nchs/data/abus/abus98.pdf
 How to Find (and Keep) the Employees You
 Want. [www.nytimes.com/indexes/2005/05/
 15/national/class/index.html](http://www.nytimes.com/indexes/2005/05/15/national/class/index.html)

Peter Muennig
Mailman School of Public Health
Columbia University
New York, USA
pm124@columbia.edu

Housing and Mortgage Market Governance

Reynold Nesiba

Introduction

The provision of housing constitutes one of the most basic and universal activities of all human societies. It is critically important for two key reasons. First, humans require shelter and housing provides this essential service. Second, investment in housing stock has become an important source of wealth accumulation for its occupant-owners as well as for real estate investors. For these reasons, governments have encouraged housing construction and its finance through tax incentives, various forms of planning and regulation, as well as through its direct provision by government agencies.

In most developed countries, the housing purchase involves a significant expense and the use of credit. For example, according to the United States Census Bureau in 2003, median household income was \$43,318. At the same time, the National Association of Realtors reports that the median home purchase price for existing single-family homes was \$170,000. Thus, for the vast majority of buyers in the US, a typical home purchase will require taking on significant debt.

To purchase a house, homebuyers around the globe use a special form of credit called a mortgage. A mortgage is a long-term loan where the property being purchased serves as security. If the buyer should default on his or her payments, the lender has the right to repossess the property. The mortgage interest rate may be fixed or vary with changes in overall market rates. According to the Mortgage Bankers Association, in March 2005, 36.6% of US mortgages have adjustable rates.

In the US, Netherlands, and Denmark a typical mortgage is for 30 years. In France, Greece, and Italy, a 15-year term is more typical. Austria, Australia, Canada, and the UK fall somewhere between the typical 15 to 30 year loan terms. In Central and South America, a history of volatile exchange and interest rates have made long term mortgages a high risk venture on the part of lenders. Therefore, they are rarely offered to potential homebuyers. For middle and upper income people in both developed and developing countries, housing is generally rented, built, or purchased directly in the market. However, it is often encouraged through various public programs and tax incentives. For lower income people in the developed world, governments have recognized the need to provide public housing for those least able to pay for it. Unfortunately for people in developing countries, little public assistance is provided to assist with housing provision.

Housing Types in Various Nations

According to a US Census Bureau estimate for July 1, 2003, the United States had 120,879,390 housing units for a population of 290,809,777 people for an average of 2.4 people per housing unit. Most (78 percent) of these units are located in cities or suburbs. About one to three million net housing units are added each year when one considers both demolition and new housing construction.

Housing units can be divided into three broad types. First, most units (76 percent) are single-family houses. These include detached houses, various forms of manufactured housing, including mobile homes or trailers, and row houses. A *row house* or *brownstone* refers to three or more adjacent homes, usually with multiple stories, that share an adjacent wall with its neighbor. A *terraced house* is a more stylish version of this structure with uniform fronts and heights. Second, 17 percent of US housing units can

be described as small (2-19 unit) apartments. These may be renter or owner-occupied. In the US a *duplex* refers to an apartment complex with two units, a *triplex* to one with three, and a *fourplex* or *quad* to a unit with four. Third, the remainder of US housing units (7 percent) consists of large apartment buildings (20 units or more), various-sized condominiums and cooperative forms of housing. These latter two housing types require additional description.

In the US and Canada a *condominium* or *townhouse* refers to an ownership arrangement where individuals own their own housing unit, but share common ownership of areas such as the facility's lobby, commons area, hallways, land, parking lot, and swimming pool. Owners typically pay a mortgage payment for the purchase and an association fee for the ongoing maintenance, repair, and insurance of the common areas. In Australia this arrangement is legally referred to as *strata title* that literally refers to units being owned at different levels. In contrast, a *cooperative* refers to a housing arrangement where residents own shares in a housing corporation. The corporation owns the land, common space, and individual units. Residents are given the right to occupy a certain apartment or unit. In most cases, condos and co-ops are located in large complexes. However, these arrangements can also be created among the residents of small apartment complexes, single-family homes, senior housing, or mobile home owners.

In the English-speaking world, people often use different words to mean roughly the same thing. A prospective tenant might look for an *apartment*, *tenement*, or *flat* in which to live. The property might be *for rent* or the prospective tenant might be looking for a place *to let*. In the US and Canada a condominium has traditionally been housing that someone buys. In other places, condos are also available to let. One is unlikely to

find a duplex in the UK. However, one might find a *2-unit rowhouse* or a *semi-detached* home, either of which mean essentially the same thing.

Housing Structure and Home Ownership

Given the highly unequal distribution of income and wealth in the world, one can also rightly expect the quality and quantity of housing to differ widely, especially between the developed and developing worlds. Combine these inequalities with differences in climate, available building materials, and cultural differences, and the result is tremendous diversity in the world's housing stock. Even among the developed countries of Europe and North America, rates of home ownership differ markedly.

According to Parsons (2003) housing construction differs greatly even among a short non-exhaustive sample of countries and regions that includes Australia, Brazil, Bangladesh, Malaysia, the Pacific Islands, Botswana, Sudan, and Mongolia.

Australia's city dwellers tend to live in detached homes made of wood or brick. These are frequently built with verandas as well as a patio or backyard for a barbecue. Roofs of red clay tile are popular, durable, and provide a distinctive aesthetic character.

In Brazil, housing for middle and upper class people is similar to those throughout Europe and the Western hemisphere. There are separate rooms for the kitchen, bedrooms, living room, bathroom, and dining room. Houses are often made of red brick or concrete block. Like Australia, red tile roofs are popular. Ceramic tile is a popular floor covering. Because of the moderate temperatures, centralized heating and cooling systems are modest—a window air conditioner or small portable heater for instance—or nonexistent. Like much of Latin America, an electric heater attached to the showerhead provides hot water, rather than a

gas-fueled water heater. Unfortunately in Brazil only about half of the population can afford houses like those described above. The World Bank suggests that the informal housing sector is growing four times as fast as the average urban growth. Housing needs are outstripping supply.

Differences in climate and proximity to the ocean create special housing needs. For instance in Bangladesh, Malaysia, and Indonesia, homes near the water are often built on stilts or embankments. This is done to protect residents from flooding and to enhance air circulation in these warm tropical environments. Rural residents of peninsular Malaysia, live in villages called *kampungs* (or *kampongs*). Their houses are built on stilts with walls and floors made of bamboo or wood. Roofs are traditionally thatched, although those with more wealth use tin or tile. In Indonesia even multi-family longhouses are sometimes built on stilts.

Among Pacific Islanders, home shape varies from island to island, but the basic framework is similar. Houses consist of wooden frames topped with thatch made from palm fronds. As corrugated steel has become less expensive and more available, it has grown in popularity as a durable alternative to thatch and fronds.

In rural southern Africa, the Tswana families of Botswana live in dwellings made up of three or four circular huts arranged around a central courtyard or *lolwapa*. Tall poles encircle the walls of the huts and support a roofing framework on which thatch is used to keep out the sun and rain. In Sudan, housing differs between the Nubian Muslims of the north and the Christian Nubians of central and south. In the Arab north, rectangular houses constructed of mud brick with flat roofs are common. In Southern Sudan, round huts with conical-shaped roofs of grass, wooden poles, and millet stalk predominate.

One of the more distinctive forms of shelter in the world is the Mongolian yurt. Mongolian herdsman require a portable structure that allows their families to move with their yaks, sheep, goats, camels, and cattle. Yurts meet this need with their circular dome-shaped frame of latticework covered by several layers of felt lashed together with rope.

In the cases above from Asia, Africa, and the Pacific Islands, there is little government involvement needed (or offered) in housing construction decisions. Individuals or small groups working together manage this social provisioning process. These descriptions should also remind us that in many developing countries, particularly in rural areas, homes are built without access to basic amenities that many of us take for granted. Piped water, sanitary sewer, electrical energy, fuel gas, telephone, cable television, and high-speed Internet connections—let alone a two-stall garage—are simply non-existent for billions of this planet's inhabitants. The nonexistent or underdeveloped infrastructure and insufficient supply of some of these services is a policy concern even in the more remote regions of the US, Canada, and Australia. This is particularly true of Native-American reservations or Aboriginal land areas.

Differences in Rates of Home Ownership

Another important difference in housing around the world is with respect to private ownership. Table 1 illustrates that even among wealthier countries, which are members of the Organization for Economic Cooperation and Development (OECD); international rates of homeownership vary from a low of 42% in Germany to a high of 85% in Spain. Data for differences in homeownership rates in developing countries is far more difficult to obtain and is often less

clear because of poorly defined property titles and less developed institutions of ownership.

Even in the US, where about 68% of families own their own homes, rates of homeownership vary significantly by race and income level. For instance, 76.2 percent of non-Hispanic whites owned their own home in the second quarter of 2004. In contrast, fewer than half of blacks (49.7%) and Hispanics (47.4%) own their own homes. In the fourth quarter of 2004, only slightly more than half (52.2%) of households with incomes below the median family income owned their own homes. For households in the top half of the income spectrum, the vast majority (84.6%) owned their own home.

Table 1. Share of Owner-Occupied Housing Among Select OECD Countries in 2002

<i>Country</i>	<i>% share</i>
Australia	70
Austria	56
Belgium	71
Canada	66
Denmark	51
Finland	58
France	55
Germany	42
Greece	83
Ireland	77
Italy	80
Japan	60
Luxembourg	70
Netherlands	53
New Zealand	65
Norway	77
Portugal	64
Spain	85
Sweden	61
United Kingdom	69
United States	68

Source: Adapted from Catte, Girouard et al (2004).

According to Forrest and Lee (2004), rates of homeownership have increased dramatically in Hong Kong despite high rates of housing price inflation in the 1980s up until the Asian financial crisis of 1997. In 1971 only 12.7

percent of residents owned their own home. This increased to 56 percent by 2002—putting its rates of homeownership above that of Germany and similar to that of France. This is striking given that Hong Kong was officially transferred from a British colony to a part of Communist China in 1997.

Public Policy Issues and Regulation

Despite the enormous diversity that exists in housing design, construction, and rates of ownership, countries share many similar issues with respect to housing and housing finance. The most basic issue for governments is ensuring that all have access to housing. In particular this includes low-income people, the elderly, minorities, refugees, and/or others who have difficulty providing basic housing services for themselves and their families. A second and closely related concern is that the habitation provided is safe, sturdy, and appropriately located. We address this second concern first. Although both developing and developed countries share these same issues, they do differ with respect to intensity of the problem as well as the resources available to address these issues.

Zoning

Some of the earliest government policy interventions regarding housing provision go back to Ancient Babylon, Greece, and Rome and were the precursors to contemporary zoning and building codes. Modern governments, typically at the local level, use zoning laws to direct the ways in which land can be developed and used. Zoning is used to distinguish where residential housing, commercial, industrial, or recreational activities can take place. In the US, local communities use zoning to regulate where different types of housing can be located. For instance, certain neighborhoods may be limited to the development of single-family

detached homes. Mobile homes, apartment buildings, duplexes, or other multiple-unit-housing would be prohibited. These other types of housing may be allowed elsewhere where they fit the zoning requirements.

The hope is that through prudent government management and oversight, new housing development as well as commercial development can proceed in a manner consistent with existing infrastructure, transportation, and pre-existing developments. In the end, zoning aims to protect property values and assure rational development processes. Unfortunately, zoning can also be used as a barrier to desegregation and to the creation of new affordable housing.

Residential Building and Construction Codes

Like zoning, building and construction codes are also a form of local government planning. These codes refer to regulations concerning the construction and occupancy of a building that are adopted and administered for the protection, health, safety, and welfare of the public. They often involve specifications about building design and materials, as well as construction standards for fire abatement, structural load, mechanical, electrical, sewage, fuel gas, and plumbing work.

In developed countries, housing that is “up to code” is said to be *standard housing*. This type of housing provides hot and cold running water, a sanitary sewage system, electrical lights, proper heating and ventilation and is otherwise safe, clean, comfortable, with enough space for all members of the household. Housing units that fail to meet local codes are viewed as *substandard housing*. This type of housing is of poor construction, unsafe, unclean, overcrowded, or otherwise deficient in some essential way. By this definition much of the housing in Africa, Asia, and Latin America is substandard. Given that 3 billion of the

Earth’s 6 billion inhabitants live on less than \$2 (US) per day, it seems plausible that most of the world’s housing is substandard. Even in the US, particularly in urban slums, Native-American reservations, and some highly segregated areas, substandard housing persists.

It is often through the adoption and enforcement of codes that governments attempt to ensure security and increase access to its housing residents. For example, in 1990 the US Congress passed the American with Disabilities Act (ADA). It mandates that public buildings be made accessible to those with disabilities. This includes ramps, wider doorways for wheelchairs, automated door openers, and other forms of equal access. Although the main aim of the legislation excludes private homes, the act has clear implications for some forms of housing. For instance, hotels, homeless shelters, nursing homes, and other forms of public residential facilities must conform to ADA standards.

Building and construction codes are also written to require builders to take into account differences in the physical environment. Housing built in earthquake zones requires different structural considerations than those located near flood plains, or in areas frequented by hurricanes and tornados. Similarly, the building needs of warmer climates differ from those of cooler climates where centralized heating, double-paned windows, and insulation are often standard. The aim is the creation of safe, comfortable, and affordable housing. Even in developed countries, this can be difficult to achieve.

Homelessness

In a famous and highly regarded two-part series in the *New York Review of Books* in 1994, sociologist Christopher Jencks reviewed much of the academic literature on homelessness in America. His popular and

persuasive analysis was developed into an influential book, *The Homeless*, in that same year.

His account focuses on understanding the most visible of the homeless—the street people and/or those who make use of emergency shelters. By drawing on an extensive review of the literature, Jencks attempts to understand the causes, extent, and possible policy solutions to homelessness.

He sees the homelessness of the 1970s and 1980s as caused by a combination of changes in personal behavior and public policy. These include 1) cuts in state support for mental health and the deinstitutionalization of the mentally ill, 2) declining rates of marriage and increasing rates of out-of-wedlock births, 3) the cheap and easy availability of crack cocaine, and 4) cuts in welfare benefits. For 1987-1988, Jencks estimated that there were about 400,000 homeless people in America—an estimate assailed as far too low by advocates for the homeless who prefer a broader definition of what constitutes homelessness.

In terms of solutions, Jencks posits that increases in welfare payments and greater access to subsidized housing for families will help address the immediate emergency need for shelter. A day-labor market supported both publicly and privately, would help ensure that those with the ability to work would be able to find employment on an ongoing but intermittent basis. This would be particularly useful to those with drug or alcohol addictions who find that occasional absences from work—because of a binge, hang over, or residential treatment experience—terminates whatever access to income and opportunity they previously enjoyed.

For the mentally ill, Jencks suggests that some may require a return to institutionalization or other assisted living arrangements. Others require better access to

social services, medication, and more careful monitoring with an aim toward maintaining their ability to live independently.

What Jencks makes very clear is that public policy solutions to address housing and mortgage market governance issues are far broader than simply providing more housing or housing finance. He shows that mental illness, addiction, broken family structures, and declining public assistance all play a role in the problems of homelessness and that a holistic policy approach is needed to address this important human concern. Nevertheless, one direct way to address a shortage of affordable housing—consistent with Jencks' proposed solutions—is simply for the government to provide it.

Public Housing and Housing Assistance

In the US, 3,300 local public *housing authorities* play an important role in assessing housing needs and responding by planning, developing, and managing area projects. These projects may entail the construction of new housing, the purchase and refurbishment of existing housing, the leasing of existing housing stock, and/or the administration of housing subsidies. In most cases, local housing authorities receive funding from the federal government through the US Department of Housing and Urban Development (HUD). HUD was created as a cabinet-level agency in 1965 and continues to be the lead federal agency on housing issues.

Like many other governments around the world, the US meets the housing needs of low-income people by providing public housing and rental subsidies. Public housing provides safe, affordable, housing to families, the elderly, and persons with disabilities who otherwise could not afford to pay market determined rental rates. The country's 1.3 million public housing sites vary from scattered site single-family homes to high-rise apartment buildings.

To qualify, low-income people must meet certain income guidelines (these vary from program to program), make their monthly rental payments, maintain the home, avoid illegal activities, and otherwise comply with the requirements of local housing authorities. It is the local housing authority that serves as the landlord to the low-income tenant.

Public housing in the US has had a checkered history. It began in earnest in the late 1930s and 1940s and is usually associated with large, high-rise towers such as the Queensbridge Houses in New York City and the Robert Taylor Homes on the Southside of Chicago. At one point, the Robert Taylor Homes consisted of 28 buildings with 16 stories each. It housed a total of 20,000 low-income residents creating one of the largest concentrations of poverty in America. It also became a haven for crime, drugs, and general lawlessness. Consensus has emerged that scattered site or mixed-income developments are better than large segregated projects. In the late 1990s, the Chicago Housing Authority began tearing down public housing high-rises including parts of the Robert Taylor Homes complex and engaging in multi-million dollar redevelopment programs to better meet community needs.

Rental assistance in the US is commonly referred to as “Section 8” or the “Housing Choice Voucher Program.” It is aimed at very low-income families and individuals—those receiving less than 50 percent of county or metropolitan area income. Despite this target audience, it was created by Congress and signed into law by President Nixon in 1974 as an inducement to private developers. The hope was that this subsidy arrangement would serve as an incentive for the construction of privately built affordable housing.

With tenant-based housing subsidies, applicants receive vouchers that can be used to rent privately owned properties. The

voucher recipient is responsible for finding his or her own housing. However, before a voucher can be paid to a landlord, the property must be inspected to ensure it meets health and safety standards. Sometimes local housing authorities will provide a list of properties that have already passed a physical inspection. Qualifying applicants may also choose a residence that has not been previously inspected. If the landlord agrees to accept a voucher, the housing authority will send out an inspector to ensure that the property qualifies before a voucher is made available.

A second type of HUD voucher, the Project-Based Rental Assistance Contract (PRAC) subsidizes building projects, as opposed to individual renters. Typically PRACs are used to encourage housing projects designed exclusively for a particular clientele, such as the chronically homeless, the mentally ill, or the elderly. For example, a project might be designed for elderly qualifying residents (who typically earn 50 percent or less of adjusted median income and be 62 years of age or older). They would be required to pay 30 percent of their income toward their housing and utilities. The balance would be subsidized by HUD’s PRAC with the project owner.

One of the ongoing problems with both of these approaches in the US is that they are very popular and oversubscribed. Because of budget constraints, popular demand cannot be met and some housing authorities have long waiting periods before newly qualified people are accepted into the program. Similarly, qualification for PRACs is through a very competitive grant award process.

The use of public housing, called “council housing” in the United Kingdom, is far more developed and has a longer history than that in the US. Housing problems were acute by the late 19th century when tenement block housing began to be provided by

philanthropists. The state became involved following passage of the Housing the Working Classes Act of 1890. Its passage ushered in the building of houses and flats by local housing councils. World War II led to the damage of almost one-third of the housing stock in the country. The state responded with centralized tower projects in urban areas as well as the development of *New Towns*. The New Towns Act in 1946 was passed to foster the development of housing units outside of center cities with an aim of reducing urban overcrowding. By the end of the 1970s, 40 percent of the UK population lived in Council Housing. During the Thatcher era, privatization was emphasized and council homes were sold to many residents on attractive terms. In addition, much of the council housing stock has been transferred to non-profit housing associations. As of 2005, 20 percent of UK residents live in housing owned by local councils or housing associations.

This is in sharp contrast to both the US and to Singapore. In the US only about 1 percent of residents live in public housing. In contrast, 85 percent of Singaporeans live in public housing. One important difference is that in Singapore, 90 percent of these “HDB flats” are owned by those living in them. They are still considered “public housing” because they are built and maintained by the Housing and Development Board. Also unlike the US, Singaporean apartment developments are designed to be self-contained with schools, supermarkets, medical clinics, and other shopping and restaurant facilities designed into centers. Ideally they are also linked to public transportation routes to facilitate commutes to work and travel throughout the area. US public housing developments have been criticized as often being too isolated from employment, shopping, and other essential facilities.

Many countries around the globe engage in some sort of public housing. However, it is often referred to in slightly different terms. In Ireland, public housing is referred to as “Local Authority Accommodation.” New Zealanders use “state housing”, while the government of Hong Kong meets housing needs through a “Home Ownership Scheme.”

Rent Control

Rent control laws of various types have been employed in many different countries and local jurisdictions around the world. One common version prohibits landlords from evicting a current tenant and limits the amount of increase in rent on that tenant. Laws differ with respect to whether landlords may raise rents when a residence is vacated. The intent of rent control is to prohibit unreasonable or excessive increases in rent with an aim of ensuring the existence of affordable rental housing. For these and other reasons, countries as diverse as India, Pakistan, Singapore, the Philippines, Ghana, South Africa, and Brazil use or have used rent control policies.

According to Walter Block (1994) rent control in the US has a history going back to its nationwide use during World War I with the intent to prevent wartime housing shortages. Controls were suspended in 1929 but employed again in 1941 during World War II. Following the war in 1948, the federal government effectively handed off jurisdiction on this issue to state and local governments. Many discontinued rent control in the 1950s. New York City is the most famous example of one city that did not. However, rent control continues in various forms in the US—in Los Angeles, San Francisco, Santa Monica, Seattle, and Washington D.C. It is also used across the Canadian border in Toronto, Ontario.

Kaushik Basu and Patrick M. Emerson (2000) suggest that World War II also served

as a spur for rent control for France, Germany and Sweden. Controls were maintained following the war to ensure that returning troops would not cause rents to spiral upward. The authors also point out that the inflation of the 1970s led the states of California, Connecticut, Massachusetts, New Jersey, and New York to implement rent controls in response.

Malpezzi and Sa-Aadua (1996) suggest that the tendency for the housing sector to experience high inflation in African countries has led many countries there to institute price controls. It is argued that since the short-run profits accrue primarily to landlords and producers, the rising prices do not necessarily stimulate long-run supply of housing. Price controls in Africa include not only rent control, but also price controls on building materials, interest rates, exchange rates, and foreign exchange rates.

In Dubai, a 2005 decree from Crown Prince Shaikh Mohammed bin Rashid al-Maktoum limits rental increases on all leased properties to a 15 percent increase. Some tenants there had reported rental increases of 100 percent in the previous year. Business leaders feared that rising costs would threaten Dubai's international competitiveness.

Some economists oppose rent controls based on a highly simplified perfect competition model of the housing market. In this textbook model, a government-regulated price reduces the incentives to build new apartments and decreases the willingness and ability of landlords to maintain their properties. This creates a shortage of housing at the rent-controlled price and thereby increases the time it takes for one to find an apartment. Despite this critique, beneficiaries of rent control argue that the benefits of rent control to lower income tenants are greater than its costs (lost rents) to wealthy landlords.

Tax Incentives: Mortgage Interest Deduction and Low Income Housing Tax Credits

In most developed countries, markets for both real estate and for home mortgage credit are well developed. One of the most common and expensive incentives is for countries to provide individual income tax deductions to encourage homeownership. According to Van den Noord and Heady (2001) many countries allow homebuyers to deduct the interest paid on their mortgages from their income taxes. These countries include Belgium, the Czech Republic, Denmark, Finland, Italy, Luxembourg, the Netherlands, Norway, Sweden, Switzerland, and the US. In the US, state and local governments further encourage "first-time-homebuyers" with programs that provide below market rates of interest as well as down payment assistance on mortgage loans.

Tax incentives are also used to encourage the creation of affordable housing. In the US, the Low-Income Housing Tax Credit (LIHTC) works by allowing the states to issue Federal tax credits to support the purchase, refurbishment, or new construction of affordable rental housing. The credits are used to offset taxes owed on other income and are often sold to outside investors to raise development funds for a particular project. To qualify for tax credits, the housing project must meet specific guidelines for the share of units set aside for low-income people and the amount of rent to be charged. With annual budget authority of about \$5 billion, the LIHTC program is the principal program for supporting the production of new and rehabilitated rental housing for low-income households. Between 1995 and 2001, the program annually created an average of 90,000 units in 1,300 projects.

Islamic Approaches to Mortgage Lending

Islamic teaching prohibits the charging or receiving of interest (riba) as a return on

capital. However, Sharia'a law does allow buyers and sellers to share the risk and rewards (e.g., profits) from investment and its financing. To comply with Sharia'a law, financial institutions in the Islamic world have created a variety of different financial agreements to facilitate housing finance. Four of the most common are *ijara*, *murabaha*, *musharaka*, and *istisna'a*. Robson (2006) says "In an *ijara* agreement, the mortgage institution buys the property and the customer makes monthly payments towards the original sale price and also pays rent for occupying the property. Ownership is transferred when the balance is paid off." *Murabaha* is a form of deferred sale. The bank purchases the property and resells it to the buyer at an agreed upon profit margin. The bank is paid through an initial down payment and installment payments over time.

Musharaka is a form of equity partnership. Here the bank buys the property and leases it to the customer who makes installment payments. With each payment the buyer's equity increases. At the end of the leasing period, the buyer owns the property. Buyers who want to build their own home use *istisna'a*, or progressive financing.

Robson notes that *ijara* mortgages dominate the Middle East region for both Muslims and non-Muslims because of its flexibility. It allows for a small down payment, a floating rate, and a time period of up to 25 years or more.

Secondary Mortgage Market

When a mortgage lender makes a loan in the primary market, they can choose to hold that loan in their portfolio or to sell it in the secondary market. By selling the loan, the lender is able to raise additional funds that can be used to fund additional mortgages. In the United States, Government Sponsored Enterprises (GSEs) such as the Federal National Mortgage Association (FNMA), The

Federal Home Loan Mortgage Corporation (FHLMC) and Governmental National Mortgage Associations (GNMA) play an important role in creating and facilitating a secondary market for mortgages.

FannieMae.com asserts that three primary benefits result from a well functioning secondary mortgage market. 1) It addresses imbalances of mortgage credit among regions of the United States by making funds available to capital-deficient areas of the country to finance new mortgage originations. 2) It allows lenders to originate mortgages for sale rather than for portfolio investment. 3) It standardizes mortgage loans, thereby attracting investors who traditionally have not invested in the primary market, further strengthening this market.

These results make the mortgage market more liquid and leaves mortgage lenders with less risk exposure than they would otherwise face. Borrowers benefit from the lower resulting interest rates and the increased availability of housing credit. Secondary mortgage markets have been slower to develop in other countries around the world. However, the International Finance Corporation, the private sector arm of the World Bank, has demonstrated its support of secondary mortgage markets by financing projects in Trinidad and Tobago, Argentina, and Mexico in an attempt to spur housing finance.

Racial Discrimination and Fair Housing

The legislative campaigns to eliminate discrimination in the housing and mortgage markets have been relatively recent. In the UK the first (1965), second (1968), and third (1976) Race Relations Acts prohibited discrimination based on race. However, unlike the grass roots community reinvestment movement that served as a motive for American legislation, in Britain the impetus for these legislative acts came

from large increases in black immigrants moving to Britain (MacEwan 1991). In South Africa, the legislation eliminating *apartheid*—the legally sanctioned system of racial segregation—as the law of the land did not occur until 1990. In the U.S., the Fair Housing Act of 1968 and Equal Credit Opportunity Act (ECOA) of 1974 were passed to forbid suppliers of housing, housing finance, and their agents from denying housing or mortgages based on an applicant's race, color, religion, national origin, age, sex, marital status or receipt of income from public assistance. The Home Mortgage Disclosure Act (HMDA) of 1975 and the Community Reinvestment Act (CRA) of 1977 (both subsequently amended) were respectively passed to increase access to bank loan records and to affirm the responsibilities banks have to local communities and individuals.

Williams *et al* (2005) suggest that the nature of racial discrimination and inequality in the mortgage market has changed. The old inequality consisted of blacks and Latinos facing lower rates of homeownership, higher rejection rates on mortgage applications, and greater degrees of racial segregation. Segregation was caused, in part, by “redlining.” This refers to the practice whereby some banks would refuse to make loans in certain areas—designated by a red line—on a city map. The new inequality deceptively appears to be an improvement. Minorities have higher rates of homeownership. In some cases they have experienced relative decreases in rejection rates. However, these gains have come with new hidden costs. Increasingly, minority borrowers now pay significantly higher interest rates and fees on their “subprime” (refers to the perceived quality of the loan) mortgages compared to comparable white borrowers and in some cases are exposed to predatory lending practices. Similarly,

minority borrowers are more frequently steered into lower-quality manufactured housing that has often been plagued by problems of construction, installation, and safety as well as higher interest and insurance costs. In short, the apparent progress of the past decade is not all that it seems. The old inequality, which denied many access to homeownership, has slowly diminished. But, for many of these homeowners, a new inequality has replaced the old. This new inequality is characterized by less desirable loan terms, exposure to predatory practices, and a lack of consumer protection.

Property Rights and Ownership Claims

Hernando de Soto (2000) has famously argued that the reason capitalism has succeeded in the west and failed in many countries, such as his native Peru, is that many non-western countries lack the legal and administrative systems essential for facilitating property claims, transfers, and financing. He argues that systemic changes require not only legal solutions, but also political and attitudinal changes to create the social institutions necessary to make capitalism work in the developing world.

It is this sort of approach that Malpezzi and Sa-Aadaw anticipated in their 1996 article on African housing policy. They assert, “urban land markets in Africa tend to be disorganized, with conflicting and often unrecorded ownership claims. Modern legal and administrative systems for surveying, recording, and transferring land titles are poorly developed.” Ownership records are often so out of date that it is impossible to determine who owns a piece of property. Without clear titles, transactions costs are very high. A buyer may have to pay multiple purported owners, mortgage financing is virtually impossible, and there is little incentive to improve land with new construction or additions when one cannot

prove ownership with certainty. These problems make administration of a property tax system difficult and expensive. Further, these authors assert that unclear property rights make illegal occupancy of property commonplace, thus allowing for slum and squatter settlements. This unplanned and unmanaged growth then results in residents living without water, sewer, or other infrastructure.

In countries or regions characterized by poorly defined institutions of private property, housing and mortgage market governance is going to be a very difficult task and piecemeal at best. It may be the case that policies aimed at “slum upgrading” rather than slum demolishing will be an improvement over past practices. By legitimizing these informal forms of housing and providing titles for their existence, inhabitants may be encouraged to see their houses not only as shelter, but also as a potential store of wealth. For much of the developing world, formalizing property title administrative systems is an important first step toward improving the provision of housing and housing finance.

Conclusion

One of the most essential and persistent international public policy issues faced by contemporary governments is to ensure that its citizens have access to affordable, safe, comfortable, housing. Over time, residential real estate has also become an important financial investment and source of wealth. For these two reasons, governments use a variety of policies and regulations to provide housing services, housing finance, and to protect the value of already existing housing stock. However, policymakers confront different challenges depending on where in the world they live.

Vast differences of climate, culture, and locally available construction materials

characterize our world, as do enormous disparities of income and wealth. Although we can categorize housing into three basic types (single-family homes, small apartments, and large apartments), the actual building materials used (bricks, wood, bamboo, mud, hides, thatch, corrugated steel), as well as the design, size, and shape differ widely. Similarly, the services that constitute a basic market basket of housing services differ widely. Where most in the developed world take clean running hot and cold water, sanitary sewer, electricity, fuel gas, and other accoutrements as standard, much of the world gets by with housing that provides little more than a thatched roof over one’s head. Where residents of wealthy countries have access to a variety of financial institutions to provide mortgages, or second mortgages, or reverse mortgages many in developing countries lack even rudimentary property rights and/or the documentation that expresses ownership, let alone mortgage credit. Poorly developed financial systems and economic volatility make access to housing finance difficult if not impossible in many countries.

These differences in concrete housing conditions as well as in the expectations of their inhabitants lead to housing and mortgage market needs and public policy responses that vary widely among countries. While the developing world strives to assist rural residents and to manage slums of squatters within and around their largest cities, the governments of the developed world use various forms of zoning and construction codes to manage and direct residential housing growth.

Among most developed countries, market mechanisms for both housing and mortgages dominate the production and distribution process. Various forms of community planning, tax incentives, and government regulation temper this free market approach. In other countries, the government takes the

lead role through direct public provision, public assistance, and various forms of rent control as well as a more limited role for markets. Regardless of the general approach or the specific policy, the primary aim is the same—safe, affordable, decent, housing for all. Once that is met, the secondary aim is to create property and financing systems that facilitate exchange and allow homeownership to also serve as a means of wealth accumulation.

Selected References

- Aalbers, M. B. (2005) "Place-based Social exclusion: Redlining in the Netherlands", *Area*, Volume 37, pp. 100-109.
- Basu, Kaushik and Patrick M. Emerson. (2000) "The Economics of Tenancy Rent Control", *Economic Journal*, Volume 110 (October), pp. 939-962.
- Best, Richard. (2005) "Successes, Failures, and Prospects for Public Housing Policy in the United Kingdom", *Housing Policy Debate*, Volume 7, Number 3, pp. 535-562.
- Block, Walter. (1994) "Rent Control: A Case Study of British Colombia", *The Mid-Atlantic Journal of Business*, Volume 30, Number 3, pp. 299-304.
- Burton, Maureen, Reynold Nesiba, and Ray Lombra. (2003) *An Introduction to Financial Markets and Institutions*. South-Western College Publishing: Cincinnati, Ohio, Chapter 12.
- Catte, Pietro, Nathalie Girouard, Robert Price and Christophe André. (2004) *Housing Markets, Wealth and the Business Cycle*. OECD Economics Department, Working Paper No. 394. Paris: OECD.
- Coccozza, Christopher R. and William J. Supper. (2006) "Reverse Mortgages: A Troubled Past, a Promising Future?", *Journal of Financial Planning*, Volume 19, Number 1, January, pp. 36-39.
- DeSoto, Hernando. (2000) *The Mystery of Capital: Why Capitalism Triumphs in the West and Fails Everywhere Else*. New York, Basic Books.
- Ferguson, Bruce. (2004) Housing Finance Options for Low and Medium Income Families: Analysis of the Latin America Experience. *Housing Finance International*, Volume 18, Number 3, March, pp. 11-14.
- Forrest, Ray, and Lee, James. (2004) "Cohort Effects, Differential Accumulation and Hong Kong's Volatile Housing Market", *Urban Studies*, Volume 41, Number 11, October, pp. 2186-2196.
- Alan Gilbert. (1994) *The Latin American City*. London, Latin American Books, & New York: Monthly Review Press.
- Gray, Jim; Jay Marcus and Jolie Marie Carey. (2005) "Cooperative Housing", *Journal of Housing and Community Development*, Volume 62, Number 6, pp. 20-24.
- Harsman, Bjorn and John M. Quigley. (1991) (Editors) *Housing Markets and Housing Institutions: An International Comparison*. Norwell, Massachusetts, Kluwer Academic Publishers.
- Jackson, Alphonso. (2006) "Tough Choices for Housing", *National Mortgage News*, Volume 30, Number 19, 13 February, pp. 4-14.
- Jencks, Christopher. (1994) *The Homeless*. Cambridge MA: Harvard University Press.
- Karakas, Cem and Onur Ozsan. (2005) "Housing Finance Practices and Development of a Secondary Mortgage Market in Turkey", *Housing Finance International*, Volume 19, Number 3, pp. 19-26.
- Malpezzi, Stephen and Sa-Aadaw, J. (1996) "What Have African Housing Policies Wrought?", *Real Estate Economics*, Volume 24, Number 2, pp. 133-160.

MacEwan, Martin. (1991) *Housing, Race and Law: The British Experience*. London and New York: Routledge.

Nesiba, Reynold F. (1999) "Housing and Mortgage Market Discrimination", in Phillip Anthony O'Hara (Editor), *Encyclopedia of Political Economy*. Volume 1. London and New York: Routledge, pp. 210–214.

Parsons, Jane. (2003) *Geography of the World*. London: Dorling Kindersley.

Porteous, David. (2005) "Setting the Context: South Africa", *Housing Finance International*, Volume 20, Number 1, pp. 34-39.

Robson, Victoria. (2006) "Opening Doors", *MEED: Middle East Economic Digest*, Volume 50, Number 4, 27 January, pp. 39-41.

Romeo, Jim. (2005) "A Roof of One's Own", *Planning*, Volume 71, Number 11, December, pp. 12-16.

Ross, Stephan L., and John Yinger. (2002) *The Color of Credit: Mortgage Discrimination, Research Methodology, and Fair-Lending Enforcement*. Cambridge MA and London, The MIT Press.

Serlen, Bruce. (2005) "Corp. Housing Players Expand Development in Asia", *Business Travel News*, Volume 21, Number 5, 29 March, pp. 42-44.

Shaw-Smith, Peter. (2006) "Affording a Home", *MEED: Middle East Economic Digest*, Volume 50, Number 4, January, pp. 44-48.

Van den Noord, Paul and Christopher Heady. (2001) *Surveillance of Tax Policies: A Synthesis of Findings in Economic Surveys*. OECD Economics Department Working Papers No. 303. Paris: OECD.

Williams, Richard, Reynold Nesiba and Eileen Diaz McConnell. (2005) "The Changing Face of Inequality in Home Mortgage Lending", *Social Problems*, Volume 52, Number 2, pp. 181–208.

Wood, David G. (2005) "Public Housing: Distressed Conditions in Developments for the Elderly and Persons with Disabilities and Strategies Used for Improvement", *GAO Reports*, 9 December.

Wood, David G. (2004) "Fair Housing: Opportunities to Improve HUD's Oversight and Management of the Enforcement Process", *GAO Reports*, 21 April.

Websites

www.bartleby.com/65/ho/housing.html This encyclopedia offers a somewhat dated, but nevertheless useful, overview of housing issues in the US, UK, and other countries.

encarta.msn.com See Microsoft ® Encarta ® Online Encyclopedia (2004) entry under "housing (shelter)" for a historical discussion of accommodation, as well as a few pictures of housing from around the world.

www.geocities.com/torontotenants/index.htm 1 Toronto Tenants provides information about rental rights, rent control, as well as a "tenants association organizing guide" for Ontario.

www.huduser.org HUD USER provides a wide variety of housing information, publications, and datasets related to the US housing market. Most of the information can be downloaded free of charge.

www.unhabitat.org "The United Nations Human Settlements Programme, UN-HABITAT, is the United Nations agency for human settlements. It is mandated by the UN General Assembly to promote socially and environmentally sustainable towns and cities with the goal of providing adequate shelter for all."

www.unece.org The United Nations Economic Commission for Europe programme on human settlements provides a variety of statistical data and country

profiles for transitional European economies.

www.housingfinance.org The International Union for Housing Finance provides information on housing finance and mortgage lending around the world in the hope of increasing homeownership rates. Their collection of reports and presentations on housing policy around the world is notable and found under the “information” link then “industry information” link.

*Reynold F. Nesiba
Department of Economics
Augustana College,
South Dakota, USA
reynold.nesiba@augie.edu*

Human Slavery

Edward J. O'Boyle

“No one shall be subjected to slavery or servitude; slavery and the slave trade shall be prohibited in all their forms”. *Universal Declaration of Human Rights*, Article 4.

Introduction

The League of Nations Slavery Convention of 1926 which sought “to prevent and suppress the slave trade” and to bring about “the complete abolition of slavery”, defines slavery as “the status or condition of a person over whom any or all of the powers attaching to the right of ownership are exercised” (League of Nations 1926:1-2). This treaty followed three other major initiatives to suppress slavery: the General Act of Berlin of 1885 which addressed slavery in the African colonies of the European states (*General Act* 1885:Article 9); the General Act and Declaration of Brussels of 1890 which dealt with the slave trade by land and sea (*Declaration* 1890:1); and the Convention of Saint-Germain-en-Laye of 1919 which affirmed the “complete suppression of all forms of slavery and the slave trade by land and by sea” (*Convention Revising* 1921:Article 11).

The Brussels Act was the first comprehensive treaty against the slave trade (Anti-Slavery International: no date:9). The United Nations approved a supplementary convention in 1956 that extended the abolition to include debt bondage, serfdom, treatment of married women or women given in marriage as their husband's or parents' property, and assignment of children to work where their labor is exploited (UN1956). Additional information on the international treaties adopted in the twentieth century relating to sexual exploitation, slavery, and trafficking is available from ECPAT

International (End Child Prostitution, Child Pornography and Trafficking in Children; see ECPAT 1996:1-22).

Bales (2002:2) most recently supplied a succinct definition of slavery: “a loss of free will and choice backed up by violence, sometimes exercised by the slaveholder, sometimes by elements of the state”. The UN Commission on Human Rights identifies several contemporary forms of slavery beyond the usual ones, including sale of children, child prostitution, child pornography, exploitation of child labor, sexual mutilation of female children, use of children in armed conflicts, traffic in persons and sale of human organs, exploitation of prostitution, and certain unspecified practices under *apartheid* and colonial regimes. Debt bondage is like traditional slavery because it is difficult to wipe out the debt which is passed on to the bonded laborer's children. Sharecropping is a common way of entering debt bondage (Office of the High Commissioner, no date:1,3).

Slavery robs its victims of their sacred dignity as human beings, a dignity which inheres in the very nature of every human being and is everyone's birthright. Slavery attacks the whole person—body, mind, and spirit—and reduces that person to an object or instrument for someone else's advantage or enrichment. Slavery subordinates one person to another, treating the core social values of freedom, equality, and community with contempt. Slavery scoffs at Kant's second imperative that no one may be used for the pleasure of another human being, no one may be reduced to instrumental value. Slavery denies the affirmation that every human being is due as a person under the commandment of love (John Paul II 1994:201).

In the following we present our remarks first on child slavery and adult slavery, then on the reasons this inhuman practice persists,

and finally on what is being done to root out this practice.

Child Slavery

The trusting, innocent, dependent nature of children, coupled with their lack of worldly experience, make them especially vulnerable to entrapment in slavery. For that reason, their enslavement is an even greater atrocity than adult enslavement. Estimating the extent of the various forms of child slavery is quite difficult not just because the practice is clandestine but also because the children are silenced by their own fear and survival needs (Office of the High Commissioner, no date:1).

ECPAT states flatly that no one knows for sure the number of children who are victims of commercial sexual exploitation worldwide (ECPAT, no date(a):1). ILO, however, states that there has been considerable progress in child labor research of late, and “the time is now ripe to update and refine the estimates” (ILO 2002:15). Worldwide there were 317 million children aged 5 to 17 engaged in some form of economic activity in 2004, including work which is permissible child labor. Among those children there were 126 million at work in hazardous circumstances (ILO 2006:2,14). Another 8.4 million children were involved in the unconditional worst forms of child labor@ including 5.7 million in forced and bonded labor, 1.8 million in prostitution and pornography, 0.6 million in illicit activities, and 0.3 million in armed conflict (ILO 2003:13-14; UNICEF 2002a:8). Extensive research between 1999 and 2001 by the Coalition to Stop the Use of Child Soldiers (CSUCS) covering 180 countries and territories revealed that (1) both boys and girls are counted among the 0.3 million fighting in more than 30 countries, and (2) hundreds of thousands of other children have been recruited into regular and para-military service, militia and other armed

groups. The youngest child soldier identified by CSUCS was a seven year old (Coalition 2001:1).

A separate estimate (to avoid the problem of double-counting) places the number of children who are trafficked for child labor at 1.2 million (ILO 2003:14). In its *State of the World's Children* report for 2002 UNICEF says that it is

“*gravely concerned* at the significant and increasing international traffic in children for the purpose of the sale of children, child prostitution, and child pornography (and) *deeply concerned* at the widespread and continuing practice of sex tourism, to which children are especially vulnerable as it directly promotes the sale of children, child prostitution and child pornography” (UNICEF 2002b:64; original emphasis).

While questioning the reliability of worldwide estimates of the number of children falling victim to commercial sexual exploitation, ECPAT at the same time asserts that reckoning the number of sexually exploited children in a specific country is much easier (ECPAT, no date(a):1). In Ghana ECPAT estimates put the number of girls, usually under age 10, who become the property of fetish priests for sexual and labor services in a religious atonement practice known as Trokosi at 4,500 (ECPAT, no date(b):3). An estimated 1 million girls work as maids in the Philippines for very low pay and long hours (Anti-Slavery International, no date:15). In Peru roughly one-half of the estimated 1.0 million adult prostitutes are actually children using false identity papers; in the United States the number of child prostitutes is put at 100,000 (Beyer 1996:32).

Adult Slavery

We have not been able to find reliable estimates of the extent of adult slavery

worldwide, and given the serious problems with estimates of child slavery (see ILO 2001:102), it is hazardous at best to arrive at an estimate for adults in slavery by subtracting the number of enslaved children from Bales' somewhat dated estimate of 27 million persons enslaved around the world (see Bales 1999:8). We therefore fall back on a variety of information sources without being able to attest to the accuracy of the information.

Slaves work in agriculture, brick making, mining or quarrying, prostitution, gem working and jewelry-making, cloth and carpet making, and domestic services. In addition, slaves clear forests, make charcoal, and work in shops. In the United States farm workers have been locked inside barracks and have labored in the fields under armed guards; enslaved women from Thailand and the Philippines have been freed from brothels in New York, Los Angeles, and Seattle (Bales 1999:1-33, 200).

The UN in 2000 reported that there were upwards of 200 million migrants worldwide of whom approximately 15 million were smuggled into the country where they presently reside (cited by Richards 2001:19). In 2001 between 0.7 and 4.0 million men, women, and children were bought, sold, transported, and held against their will in a form of slavery which is known as "trafficking" (U.S. Department of State 2002:1). Trafficking is the

"recruitment, transportation, transfer, harbouring, or receipt of persons, by means of threat or use of force or other forms of coercion, of abduction, of fraud, of deception, of the abuse of power or of a position of vulnerability or of the giving or receiving of payments or benefits to achieve the consent of the person having control over another person, for the purpose of exploitation (which includes)...

prostitution ... or other forms of sexual exploitation, forced labour or services, slavery or (similar) practices...servitude or removal of organs" (UN General Assembly 2001:32).

Smuggling is "the "procurement ... to obtain ... a financial or other material benefit, of illegal entry of a person into a State Party of which the person is not a national or permanent resident" (UN General Assembly 2001:41). An estimated 0.5 million women are trafficked into western Europe every year (UNICEF 2002c:4). About 50,000 women and children are trafficked in the U.S. every year (Anti-Slavery International 2002:124) for the purpose of prostitution, stripping/sexual touching, sweated labor, agricultural slave labor, domestic and other forms of servitude (Richard 2000:50). The annual profits derived from trafficking range from \$3 to \$10 billion (Schloenhardt 1999:23).

Three other bits of information from two especially noteworthy sources are compelling enough to attach to the end of this section on adult slavery. First, ILO reported that "*millions*" of persons throughout South Asia and Central and South America presently live and work under conditions of debt bondage (ILO 2001:vii; emphasis added). Second, the UN Working Group on Contemporary Forms of Slavery stated that there are known instances in which the bodily organs of executed prisoners are being exploited for commercial purposes, and UN personnel notably peacekeeping forces engage in sexual and other kinds of exploitative conduct (UN 2002:2). Third, corrupt public officials at times actually facilitate trafficking and smuggling (Richard 2000:8,15; UN 2002:1; Bales 1999:245).

Why Slavery Persists

Just as there are two principal parties to the practice of slavery -- the person enslaved and

one who enslaves -- there are two sets of reasons as to why the practice persists. On the part of the person enslaved there is a material need grounded in the unrelieved poverty and dearth of opportunities of that person=s pre-enslavement circumstances, though the linkage between poverty and slavery is neither complete nor direct (see, for example, U.S. Department of State 2002:1-2; ILO 2001:101; ILO 2002:xii; Strandberg 1999:7). Unmet need pushes that person -- if a child, his/her family or guardian may use that unmet need to push that child -- into labor which through deception, force, and violence is exploited, where the poverty continues and a form of bondage may be imposed. Additionally, there is a fundamental human need for work as such which meets the need for belonging and the need to engage in work which is challenging and creative, allowing that human being to develop more fully towards his/her full potential as a person (David 2000:3). The need to belong can be denied effectively by several means including language barrier, physical confinement, and passport seizure. The need for creative work opportunities makes the innocent and naive vulnerable to being duped and deceived by the promises of the agents of slavery (Ryf 2002:49-51). For example, the promise of marriage may entrap a girl into forced prostitution; the enticement of learning a skill or trade may ensnare a boy into domestic servitude (U.S. Department of State 2002:1).

We prefer this framework for addressing the persistence of slavery rather than the more conventional supply/demand or push/pull model because at the very core of slavery, as mentioned previously, is a devaluing of human beings which strips them of their inherent dignity as persons and reduces them to instruments for the illicit and unjust enrichment of others. Bales reports that in Pakistan, India, Mauritania, and Brazil nearly every slaveholder he met and interviewed

regarded himself as a businessman. Indeed these agents of slavery were family men and pillars of the community (Bales 2002:4). This objectification of human beings is best described in the "sex object" language routinely used to characterize the prostitute. Objectification, however, is even more common than what is represented by the practice of human enslavement. Millions of other humans are reduced to objects as John Paul II warns in *Evangelium Vitae* through murder, genocide, abortion, euthanasia, willful self-destruction, mutilation, torments inflicted on body or mind, attempts to coerce the will itself, subhuman living conditions, arbitrary imprisonment, deportation, slavery, prostitution, the selling of women and children, disgraceful working conditions (John Paul 1995:14).

Liberation and Rehabilitation

Liberating and rehabilitating children and adults who are held in slavery begin with the recognition that slavery today is a worldwide human tragedy which ultimately reduces to one human being treating another as an exploitable property rather than a human person (UN General Assembly 2001:32,41). Ryf asserts that governments and law enforcement agencies worldwide are contributing to the spread of trafficking due to a failure to recognize the problem, to outlaw the practice, and to appropriate the necessary funds to enforce anti-trafficking laws. In the United States, \$95 million was appropriated for 2001-2002 to combat trafficking but Ryf states that even these resources may not reduce world trafficking to any significant degree (Ryf 2002:69-70). Bales asserts that programs of liberation and rehabilitation are still in their infancy and no systematic evaluation is presently available. A further complication is that there has been no in-depth social science study of the

relationship between the master and the slave (Bales 2002:5).

Nevertheless the practice of enslavement is so widespread today that we cannot use our own ignorance as justification for inaction. For that reason we turn to four international agencies on how to address the problem of human enslavement. Those agencies are Anti-Slavery International, ILO, ECPAT International, and the Office of the United Nations Commissioner for Human Rights. In the following we present selectively those recommendations which have some specificity and which therefore make clear *what* should be done and *who* is to do it.

Based on its own studies, Anti-Slavery International has advanced 45 recommendations relating to government responsibilities in dealing with trafficking. Because there is no prioritization of the 45 proposals, we have selected one recommendation from five of the nine sets of recommendations in order to suggest the scope of governmental action required to reduce trafficking. (1) Persons who have been trafficked should not be prosecuted for acts such as prostitution which were performed while they were being trafficked. (2) Persons who have been trafficked should be informed of their right to asylum and be granted asylum whenever appropriate. (3) The state should provide shelters for persons who have been trafficked. (4) Laws should be enacted which allow confiscation of the assets of traffickers and use of the proceeds of the liquidated assets as compensation for persons who have been trafficked. (5) The state should not force the return of a trafficked person to his/her country of origin when there is evidence that the person may be subject to discrimination, stigmatization, or reprisal (Anti-Slavery International 2002:5-12).

The ILO recommends microfinance and microcredit arrangements which target families at risk of falling into enslavement

and which focus especially on women who are key to reducing the number of children who are trafficked. Stronger preventive labor inspection measures, the ILO argues, likely contribute to the elimination of forced labor (ILO 2001:x,102).

ECPAT International has issued a report with numerous recommended actions to protect children. Three have been selected for their specificity. (1) The state should provide a guardian *ad litem* to assist a child who is a witness in criminal proceedings which involve allegations of sexual exploitation of children. (2) Telephone help lines should be made available to children seeking assistance because they have been abused or exploited. (3) An ombudsman, institution, or agency should be appointed to hear and act on complaints from children (ECPAT 1996:15, 18, 22).

The Office of the UN High Commissioner for Human Rights has prepared a lengthy list of recommendations many of which are general. Four are selected because they are specific. (1) The state should review its legislation regarding use of the Internet for the purpose of trafficking, prostitution and sexual exploitation of women and children and enact new legislation as required to prevent such abuses. (2) The state should implement measures to prevent and sanction the confiscation of the passports of migrant workers. (3) No girl of primary school age should be employed as a domestic. (4) In depth investigations should be conducted to determine the role of corruption and international debt in fostering slavery (Office of the High Commissioner 1999:3-4).

Slavery in the end is rooted in a culture of death and despair. It will persist as long as humankind clings to those cultural values. It will not be wiped out until humankind embraces the counter-cultural values of life and hope.

Selected References

- Anti-Slavery International. (no date) *History of Anti-Slavery International*.
www.antislavery.org/homepage/antislavery/history.pdf
- Anti-Slavery International. (2002) *Human Traffic, Human Rights: Redefining Victim Protection*.
www.antislavery.org/homepage/resources/humantraffichumanrights.htm
- Bales, Kevin. (1999) *Disposable People: New Slavery in the Global Economy*. Berkeley: University of California Press.
- Bales, Kevin (2002) "The Social Psychology of Modern Slavery", *Scientific American*, 286, 4, 80-88.
- Beyer, Dorianne. (1996) "Child Prostitution in Latin America", in *Forced Labor: The Prostitution of Children*, Washington, D.C.: U.S. Department of Labor.
- Coalition to Stop the Use of Child Soldiers. (2001) *Child Soldiers Global Report*. New York.
- Convention Revising the General Act of Berlin, February 26, 1885, and the General Act and Declaration of Brussels, July 2, 1890*, Signed at Saint-Germain-en-Laye, 10 September 1919. In *Supplement to the American Journal of International Law*, Volume 15.
- David, Fiona. (2000) *People Smuggling in Global Perspective*. Paper presented at the Transnational Crime Conference, Canberra, Australia, March 9-10.
www.aic.gov.au/conferences/transnational/david.html
- Declaration of the General Act of the Brussels Conference, July 2, 1890*.
- ECPAT International. (no date a). *How Many Children are Victims?*
www.ecpat.net/eng/CSEC/faq/faq8.asp
- ECPAT International. (1996) *The International Legal Framework and Current National Legislative and Enforcement Responses*. Paper presented at World Congress Against the Commercial Sexual Exploitation of Children, United States Embassy, Stockholm, August 27-31.
- ECPAT International. (no date b) "What Makes Children Vulnerable to Sexual Exploitation?"
www.ecpat.net/eng/CSEC/faq/faq9.asp
- General Act of the Berlin Conference*. (1885)
en.wikipedia.org/wiki/Conference_of_Berlin
- International Labour Organization (2002) *A Future Without Child Labour*. Geneva.
- International Labour Organization. (2006) *Global Child Labour Trends 2000 to 2004*. Geneva.
- International Labour Organization. (2003) *IPEC Action Against Child Labour: Highlights 2002*. Geneva.
- International Labour Organization. (2001) *Stopping Forced Labour*. Geneva.
- John Paul II. (1994) *Crossing the Threshold of Hope*. New York: Alfred A. Knopf.
- John Paul II. (1995) *Evangelium Vitae*. Boston: Pauline Books and Media.
- League of Nations. (1926) *Slavery Convention*. Geneva.
- Office of the United Nations High Commissioner for Human Rights. (1999) *Report of the Working Group on Contemporary Forms of Slavery*.
www.uri.edu/artsci/wms/hughes/wgcf99.htm
- Office of the United Nations High Commissioner for Human Rights. (no date) *Fact Sheet No. 14, Contemporary Forms of Slavery*.
www.unhchr.ch/html/menu6/2/fs14.htm
- Richard, Amy O'Neill. (2000) *International Trafficking in Women to the United States: A Contemporary Manifestation of Slavery and Organized Crime*. Washington, D.C.: U.S. Department of State, Bureau of Intelligence and Research.
- Richards, Lenore. (2001) "Trafficking in Misery: Human Migrant Smuggling and

Organized Crime”, *Gazette*, Volume 63, Number 3:19-23.

Ryf, Kara C. (2002) “The First Modern Anti-Slavery Law: The Trafficking Victims Protection Act of 2000”, *Case Western Reserve Journal of International Law*, 34, 45, 45-71.

Schoenhardt, Andreas. (1999) *Organized Crime and the Business of Migrant Trafficking: An Economic Analysis*. Paper presented to the Australian Institute of Criminology, November 10.

Strandberg, Nina. (1999) *What Is Trafficking in Women and What Can Be Done?* Stockholm: Kvinnoforum/Foundation of Women’s Forum.

UNICEF. (2002a) *Adult Wars, Child Soldiers*. New York.

UNICEF. (2002b) *The State of the World’s Children 2002*. New York.

UNICEF. (2002c) *Trafficking in Human Beings in Southeastern Europe, 2002*. New York.

United Nations. (2002) *Contemporary Forms of Slavery: Sub-Commission on the Promotion and Protection of Human Rights*. Report of the Working Group on Contemporary Forms of Slavery. New York.

United Nations (1956) *Supplementary Convention on the Abolition of Slavery, the Slave Trade, and Institutions and Practices Similar to Slavery*. New York.

United Nations General Assembly. (2001) “United Nations Convention Against Transnational Organized Crime”, Resolution 55/25. New York.

United States Department of State. (2002) *Victims of Trafficking and Violence Protection Act 2000. Trafficking in Persons Report*. June. Washington DC.

Universal Declaration of Human Rights. (1948) www.un.org/Overview/rights.html

Edward J. O’Boyle
Mayo Research Institute
West Monroe, Louisiana, USA
edoboyle@earthlink.net

Inequality and Distribution

Charles M.A. Clark

Introduction

“How a society divides its social product?” is one of the three primary questions in economics (what to produce? and how to produce? are the other two). It is never a neutral question as it affects not only the well-being of the individual members of the society (their respective shares) but also the future well-being of the society. The minimum requirement for any distribution of income is that the vast majority of the members of a society receive enough to stay alive and productive. Moreover, larger shares can be used as incentives for various types of activities which the society especially needs or values. While income most directly influences a person’s well-being in the short run, wealth is an important determinant of long term economic security, thus its distribution is significant. Furthermore, the distribution of wealth involves the distribution of the ownership of society’s assets (particularly its productive assets), and therefore has political and social importance. In this entry we will look at the extent of the inequality of wealth and income; the theories used to explain levels of inequality; the role of government policy in promoting or reducing inequality; and lastly, how inequality influences economic outcomes.

Definitions of Wealth and Income

The words “Wealth” and “Income” have had, and continue to have, multiple meanings both in common parlance and theoretical analysis. Derived from the old English word “welde”, the multiple meanings of wealth range from spiritual and material well-being to an abundance of material possessions and riches. Adam Smith defined real wealth as “the annual produce of the land and labour of the

society” (Smith 1976b:12), following in this tradition of equating wealth with abundance. With the onset of neoclassical economic theory, most economists have adopted a scarcity based understanding of wealth—the ownership of scarce assets that yield a return or that can be turned into purchasing power. The word income has a similar spiritual origin, referring to the coming in of divine influence. Eventually the meaning of the word “income” went from “entering in” to “the fee paid to entering” to “the payment for some service.”

To date, there is no commonly accepted definition of either wealth or income which is both theoretically consistent with neoclassical economic theory and useable for empirical measurement and analysis. This is not for want of trying. Defining wealth and income was a key question for late 19th and early 20th century economists, culminating in John Hicks (1939,1942) and Nicholas Kaldor’s (1955) work in the mid 20th century. However, the more important issue for public policy quickly became the development of systems of national income accounting. Attempts to treat income as consumption of utility (which left one as well off at the end of the time period as one was at the beginning) and wealth as deferred consumption (savings), offered neither significant theoretical insights nor practical policy guidance. One conclusion of the attempts to define wealth and income in the neoclassical tradition is the contention that income is a flow variable and that wealth is a stock variable. Income is what is available for consumption and wealth consists of those assets which are not being consumed (with the assumption that they are being used to promote future production). Additionally, when these assets (wealth) are sold for money that will go for consumption, income is increased. And when income is spent on an

asset that does not get completely consumed in the short run it becomes wealth.

The “wealth” of a nation is composed of four factors: the size and skill level of its population (paid workers and non-paid worker all contribute to output); its capital stock; the availability of resources; and the state of technological knowledge; all of which are not easily measured (Veblen 1908). Coupled with the influence of culture, these determine the level of production of goods and services. When we measure wealth we are in fact measuring the valuations of some of these factors, valuations which reflect many factors besides productivity, and which are notoriously imprecise, both theoretically (as exposed in the Cambridge Capital Controversies) and empirically. Market valuations of these assets are often driven by expectations (such as the “dot com” stocks in the late 1990s) and manipulations (companies purchasing their own stocks) and thus are not reliable indices of their real worth (at least in the short run). Furthermore, all measures of “wealth” do not include the first factor we listed, workers (in the broadest sense of the word), who account for the most important contributions to the production process.

For the most part, economists have adopted definitions of income and wealth that governments developed for the purpose of tax assessment and collection. Personal or household income, which is usually what is important for discussions of income inequality, is defined as money that is received to support the persons current well-being, and typically includes: labor income (wages and benefits); proprietors earnings; rental income; dividends; interest income and transfer payments (Wolff 1997). Personal or household marketable wealth typically includes: owner-occupied housing and other real estate and land; consumer durables; financial assets such as checking and savings accounts, cash and currency, bonds and

financial securities and life insurance cash surrender value; equities (stocks, mutual funds, unincorporated business equity and trust equity) and miscellaneous assets such as money owed by family business or friends (Wolff 1997). Other measures make adjustments for how liquid the assets are, thus leading to more narrow definitions of household wealth (fungible wealth and financial wealth), yet one can take a broader conception of wealth (based on what wealth does in terms of providing financial security) and include social security and pension wealth (Augmented Wealth) which can increase the total wealth by 50% or more (Wolff 1997).

Trends in Income and Wealth Inequality

The Italian economist Vilfredo Pareto (1896) suggested that there is a uniform distribution of income across countries. The evidence, both over time within countries, and comparisons between countries, suggests otherwise. In Table 1 below we see the level of income inequality in various advanced capitalist countries at the end of the 20th Century, as measured by the Gini coefficient and the 90/10 percentile ratio. The Gini coefficient, which measures the difference between a country’s actual distribution of income and what a perfectly equal distribution (based on percent of population receiving an equal share of aggregate income) is one of the most widely used measure of inequality. The lower the Gini coefficient the more equal the distribution of income, with perfect equality equaling 0.0 and perfect inequality equaling 1.0. The 90/10 Percentile Ratio is easier to calculate and more intuitive to interpret, is based on a ranking of the population by income and then dividing the income share of the 90th percentile by the 10th percentile. First, Table One shows us that there are considerable differences in the levels of income inequality among the 15

countries listed. Second, the table shows that inequality rankings are sensitive to choice of measurement, though the general picture that emerges is the same. This general picture is that indicates that how one measures inequality will influence where a specific country is ranked.

Table 1 -- Income Inequality, Various Countries, End of 20th Century.

Country	Gini	Rank	Percentile Ratio (90/10)	Rank
Finland 2000	.247	1	2.90	2
Netherlands 1999	.248	2	2.98	4
Belgium 1997	.250	3	3.19	6
Norway 2000	.251	4	2.80	1
Sweden 2000	.252	5	2.96	3
Denmark 1997	.257	6	3.15	5
Germany 2000	.264	7	3.29	7
Austria 1997	.266	8	3.37	8
France 1994	.288	9	3.54	9
Canada 2000	.302	10	3.95	10
Australia 1994	.311	11	4.33	11
Ireland 2000	.323	12	4.56	13
Italy 2000	.333	13	4.48	12
UK 1999	.345	14	4.58	14
USA 2000	.372	15	5.45	15

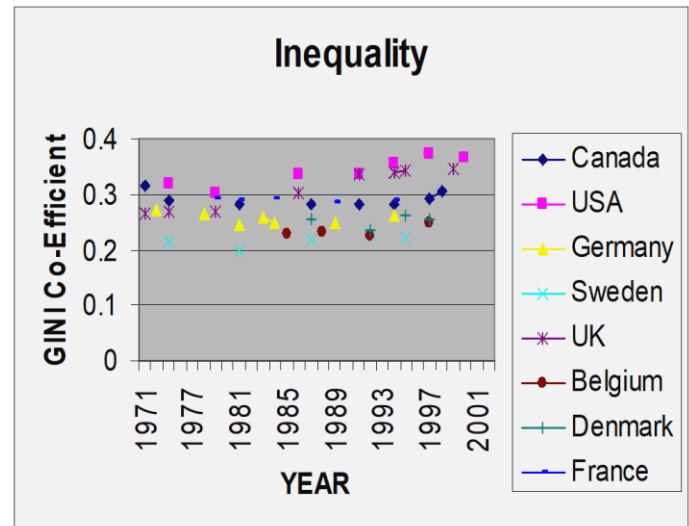
Source: Adapted from the Luxembourg Income Study. www.lisproject.org/keyfigures/ineqtable.htm.

This table also shows how important culture is in determining income inequality, with the Northern European countries generally having lower levels of income inequality, whereas the highest inequality levels tend to be in English speaking countries (UK and its former colonies). It is often suggested that their common legal traditions and attitudes toward government account for some of the high levels inequality in these countries.

A look at income inequality within a country over time also shows the possible variations. In Graph 1 we see changes in the Gini coefficient for various advanced capitalist countries from the early 1970s to

2000, which shows that while there has been a trend in inequality increasing since the 1980s, many countries have resisted this trend.

Graph 1, Gini Coefficient, Various Nations



Source: Adapted from the Luxembourg Income Study. www.lisproject.org/keyfigures/ineqtable.htm.

One of the first to become aware of the variations in income inequality was Simon Kuznets (1955), who noticed an inverted U-shaped relationship between the level of national income and income inequality. Income inequality was low for low income countries, and tended to rise as income levels rose. However, eventually the relationship reversed, with all the advanced, industrialized countries being on the downward portion of the inverted U curve. The last two decades of the 20th century bucked this trend, with inequality rising in many advanced industrial economies, and with stagnation in the trend towards greater equality in the other advanced industrialized countries. This rise in inequality was one of the major topics of research and discussion over the past twenty years, with economists blaming everything from international trade and globalization to demographic changes and the impact of the computers. The fact that some countries have had a rise in inequality while others have not experienced an increase suggests that

government policy plays a role in determining income distribution. We return to this question in the Governance Issues section of this entry.

World income inequality is a growing topic on interest among economists. While comparisons of per capita income levels between countries does demonstrate the high level of income inequality in the world, even when adjusted for purchasing parity, it relies on national averages, thus does not adjust for inequality within countries. Some recent studies have used the growing availability of household surveys to measure world inequality (Milanovic 2002; see Lundberg & Milanovic 2000 for an overview). Consistent with the methodological individualism that underlies neoclassical economic theory, these studies attempt to calculate a World Gini Index. However, the most important determinant of one's income level is where one is born and it is national factors and not individual factors that determine world income inequality levels.

Distribution of Wealth

Analysis of the distribution of wealth has always been more difficult than that of income, for the simple reason that the data has been more difficult to come by (partly stemming from the difficulties of defining and measuring wealth mentioned above). We do know that the distribution of wealth is significantly more unequal than that of income and that since wealth often yields an income, this contributes to income inequality. Table 2 shows the trends in wealth inequality in the UK and USA throughout the 20th century we see that for most of the century wealth has tended towards greater equality, yet after the mid-1980's the trend has been towards greater concentration, thus following the trend in income inequality for these two countries.

Measuring the inequality of wealth between countries is even more problematic than measuring it within a country, especially when it comes to individual wealth.

Table 2. Wealth Inequality in UK and USA, 1911-1998

Year	USA		UK	
	Top 1%	Top 10%	Top 1%	Top 10%
1911	NA	NA	70	NA
1962	33.4	67.0	33	69
1983	33.8	68.2	20	52
1989	37.4	70.5	17	48
1992	37.2	71.8	18	50
1995	37.6	71.6	19	50
1998	38.1	70.9	23	55

Source: Adapted from: USA, Wolff (2002); UK, Paxton (2002).

The World Wealth Reports, which tracks the financial status of High Net Worth Individuals (multimillionaires), gives an indication how individual wealth is distributed across the globe. In Table 3 we see how that wealth is concentrated in North American and Europe (29.5% and 30.2% of the 2003 total, respectively).

Table 3. HNWI Wealth by Region, 1997-2003 (\$ Trillion)

Region	1997	1999	2000	2001	2002	2003
North America	5.9	8.1	7.5	7.6	7.4	8.5
Europe	5.3	7.3	8.4	8.2	8.4	8.7
Asia	4.0	5.4	4.8	5.3	5.9	6.5
Latin America	2.5	3.1	3.2	3.5	3.6	3.7
Middle East	0.9	1.1	1.0	0.8	0.8	0.8
Africa	0.5	0.5	0.6	0.6	0.6	0.6
Total	19.1	25.5	25.5	26.0	26.7	28.8

Source: Adapted from the World Wealth Report 2000; 2002 and 2004.

In Table 4 we see the value of stock exchanges by region. Stocks are an important form of wealth for the stock market is supposed to provide market discipline over the operation of corporations.

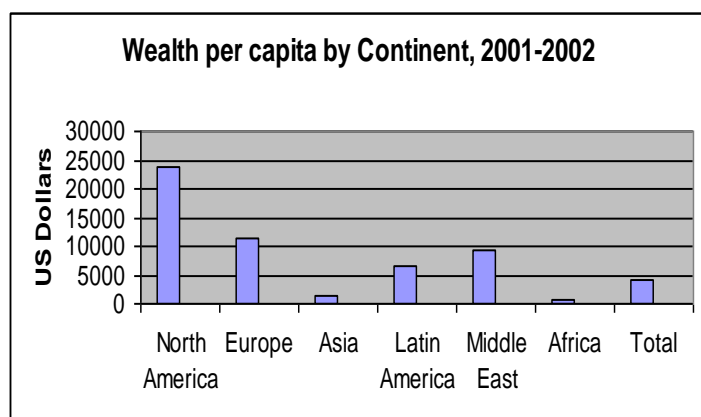
Table 4. Global Market Capitalization by Region, 1998-2002 (\$Trillion)

Region	1998	1999	2000	2001	2002
North America	14.1	18.3	16.0	14.7	11.6
Europe	8.4	10.6	9.3	7.5	6.3
Asia	5.7	9.6	4.9	4.5	4.3
ROW	0.7	1.1	0.7	1.1	0.9
Worldwide	28.9	39.6	30.9	27.8	23.1

Source: Adapted from the World Wealth Report 2000, 2002 & 2004.

In Graph 2, which presents a measure of wealth per capita by continent (keeping in mind that it is only the wealth of HNWI, however, this accounts for the vast majority of the worlds wealth), we see that world distribution of wealth is highly unequal.

Graph 2. Per Capita Wealth. Continents. 2001-2002



Source: Adapted from the World Wealth Report, 2002, and author's calculations.

Theories of Inequality

Adam Smith followed two tracks when analyzing the distribution of income. On the one hand, Smith (1976b [1776], Book One, Chapter 6) explain the distribution of income in terms of the three social classes that existed in the late 18th century: Landlords, Workers and Masters (capitalist); from whence we get the three factors of production: land, labor and capital. Underlying Smith's analysis is the assumption that market forces would generate equality. Smith noted in *The Theory of Moral Sentiments*, that "[Men] are led by

an invisible hand to make nearly the same distribution of the necessities of life, which would have been made, had the earth been divided into equal portions among all its inhabitants, and thus without intending it, without knowing it, advance the interests of the society" (Smith 1976a:186). However, on the other hand, Smith looks at the factors that produce inequality within groupings (1976b: Chapters 10, 11), which result from two types of factors: inequalities due to the nature of employments and inequalities created by government policy. From these two explanations of distribution we get the two main approaches to explaining inequality: the structural approach which looks to structural elements in society and the economy to explain inequality; and the individualistic approach which looks to individuals and their choices to explain inequality. The Institutionalist approach attempts to include both structure and agency and thus is following Smith in this respect.

Structural Inequality

When Adam Smith argued that the differences between the philosopher and the porter were not that great and due to education, he was expressing a firm belief of many of the Enlightenment philosophers of the underlying equality of men and of the role of institutions in creating inequality. Smith's class-based approach was followed by Ricardo and Malthus, but where Smith saw a "harmony of interests" they saw conflict. Ricardo's "iron law of wages" stated that workers wages would, in the long run, tend toward the subsistence level. According to Ricardo the income of the other two classes (Landlords and Capitalists) would be determined mostly by the level of population. Building on Ricardo's labor theory of value, Karl Marx developed a theory of inequality that went to the heart of the structure of capitalist society. Marx's analysis

demonstrated that the source of profits and rents was the social surplus and that the social surplus came from the fact that workers were exploited (they got paid less than the value they created) due to the structural features of capitalist societies (private property in the means of production) and unequal power relations. Post Keynesian theories of income distribution (Kaldor 1956, Galbraith 1998) are also within the structuralist tradition, as they are based on structural factors such as the degree of monopoly, the level and use of property income and macroeconomic factors like the unemployment rate.

Much of the research within the structuralist tradition is being done by Radical/Marxist and Feminist economists, emphasizing the role of discrimination and exclusion as the causes of poverty and inequality. This is seen in much of the work on gender and racial differences in incomes within capitalist societies, and in the uneven terms of trade between the rich countries and the third world. The empirical work in the structuralist tradition often emphasizes regional and industrial factors in determining incomes and thus inequality. Structuralists also often note the disparities between the education received by different social classes and that this education opportunity inequality perpetuates and increases income inequality.

Individualistic Theories of Inequality

Most current work on income inequality follows the neoclassical approach of explaining incomes just as they explain other prices, via supply and demand. Whereas Adam Smith emphasized the differences between different occupations and industries, current individualistic theories emphasize differences between individuals. Two key aspects of this approach are its rigid adherence to methodological individualism (which forces them to exclude all social and historical factors from theoretical

consideration) and its assertion that incomes are determined by productivities. As John Bates Clark, the originator of marginal productivity theory, stated: “the distribution of the income of society is controlled by a natural law, and that this law, if it worked without friction, would give every agent of production the amount of wealth which that agent creates” (Clark 1965:v).

Individualistic theories of inequality are often called “achievement” theories since they attempt to link incomes to activities or characteristics of individuals which merit remunerated. Thus the neoclassical theory of income distribution becomes a question of income determination. Underlying this approach are the two principles that 1) “all agents in the economy (i.e. individuals, firms, unions, governments) maximize a well-defined objective function; and 2) there exists a market equilibrium which balances the conflicting goals of the various players in the labor market” (Borjas 1988:21). Thus market outcomes, including income distribution, are the result of individual choices, mediated through the impersonal market. Incomes are prices, thus if they differ it is because of some supply or demand factor, and if the differences are long term, it is because of a difference in productivity.

One important aspect of this approach to explaining income inequality is human capital theory. The underlying idea behind human capital theory is that one of the most important choices individuals make in the market is the decision to accumulate human capital. The basic idea is that individuals can choose to forgo current income and consumption and instead invest their time and money into acquiring education and skills, with the pay-off being higher future incomes. Similarly, profit, rent and interest income is explained based on decisions to forgo consumption and save and invest, with market forces determining the returns on

these investments based on their productivity (how much they contribute to a firm's profitability). The empirical work in neoclassical economic theory in income inequality typically looks at how variations in various individual characteristics (such as education attainment and work experience) are correlated with variations in wage rates or incomes. In effect, each of these individual characteristics is given a price, with the assumption that these prices are determined in competitive markets. Often, factors like race, gender and union membership are included as if they were individual characteristics whose price was determined in competitive markets. During the discussion on the causes in the rise in income inequality in the last two decades of the 20th Century much of the empirical work centered on the increase in the education premium, specifically it was argued that "new economy" required greater levels of education and computer literacy and that these skills became more in demand, thus causing the wage differential for those with such skills to increase.

Some recent work within the neoclassical tradition has included structural (institutional) factors in explaining wage inequality (Blau and Khan 1998), recognizing that incomes as prices are not always determined in markets that are perfectly competitive. Moreover, other market failures are being recognized as playing a role in income determination, such as incomplete markets, imperfect information and externalities. Add to this the obvious role of institutions such as unions and collective bargaining, as well as government policies such as minimum wage and the tax treatment for different types of incomes, and you start to get a convergence between the individualistic and structuralist approaches. Another example of neoclassical theory including institutions in their explanations of income inequality can be found in the Public Choice and Rent Seeking literature

(Buchanan, Tollison, Tullock 1980; Medema 1991). While the development of the theory of rent seeking was for the purpose of arguing that government activity in the economy (besides protecting property rights) was inherently wasteful, the theory can be applied to explaining persistent inequalities in incomes. At its most basic level the theory of rent seeking states that governments create rents (incomes higher than what a perfectly competitive market would yield) thus influencing the distribution of incomes (higher incomes for those who receive the rents, lower incomes for those who either pay higher prices or higher taxes). Individuals and groups expend resources in order to obtain government policies that will transfer wealth in their direction, or to maintain such policies. At the surface it looks like this theory is a step in the direction of the long held Institutionalist argument of the connection between power and wealth, yet the key difference is the rent seeking theory has an underlying assumption that there is, at least theoretically, a régime of property rights (that somehow aren't created by government) that are neutral and thus do not produce any rents (waste). Where such a regime to exist the invisible hand of the market would dissipate all market power, and remaining inequalities would be achievement based.

Institutionalist Theories of Inequality

The Institutionalist approach to explaining wealth and income inequality takes its inspiration from John Stuart Mill's argument that the distribution of wealth "is a matter of human institutions solely. ... [I]n the social state, in every state except total solitude, any disposal whatever ... can only take place by the consent of society, or rather of those who dispose of its active force. ... The distribution of wealth, therefore, depends on the laws and customs of society. The rules by which it is determined are what the opinions and feelings

of the ruling portion of the community make them, and are very different in different ages and countries; and might be still more different, if mankind so chose" (Mill 1987: 201).

Like structural theories, Institutionalists emphasize the role of rules, culture, government policy and power in determining incomes, but they also include an active role for understanding behavior and free action (human agency), which are the central concern for individualistic theories. The Institutional theory of inequality starts with the work of Veblen, specifically his *The Theory of the Leisure Class* (1899) which takes an historical look at how inequality has been established socially, politically, economically and culturally, and the *Theory of the Business Enterprise* (1904) which looks at the large corporation as an institution for generating inequality working through the economy and the use of power in markets.

William Dugger (1987) presented an institutionalist theory of inequality which emphasizes that the distribution of income is created by the complex interaction of many factors, economic and non-economic. Dugger differentiates three such modes for distributing incomes: Industry; Hierarchy and Market. Each mode has both functional and discretionary aspects as to how income is distributed and that in an affluent society, discretionary plays the more important role. However, the instituting of power in all three modes is what is most important, influencing not only standard economic variables like monopoly but also influencing attitudes and preferences, thus individual economic actions. Furthermore, Dugger notes that inequality is not an equilibrium state, that each mode of inequality help to generate a process of either greater or lesser inequality.

The Institutional tradition has also emphasized the important role of government policy in influencing income inequality

(Clark 1996), yet it does not look at government policy as a market failure, but as necessary for the functioning of markets. That is, it recognizes that government policy could promote greater equality (such as universal education) or greater inequality (high interest rates and restrictive macro-economic policies that increase unemployment, or anti-union legislation that weakens workers bargaining power) and that the individual activities (choices) made in the marketplace are not merely economic responses to price signals, but are fully social actions, reflecting prejudices and attitudes that can perpetuate and exasperate income inequality. Economic choices are always made in a social and historical context, which influences not only the preferences, but more importantly, determine the range of options the individual can chose from.

Political, Social and Economic Impacts

That wealth and income inequality have important social and political consequences is most likely obvious to all and requires little elaboration. Wealth and incomes are an important determinant of social class and social status, especially in the highest stages of capitalist development. Other factors, such as education, are primarily a function of wealth and income. Whether one has a political voice in a society is also greatly influenced by wealth and income (Phillips 2002). Physical and mental health has been shown to be influenced by the level of income inequality (Wilkinson 1996).

While the conventional wisdom holds that inequality promotes economic growth, much of the recent research by the World Bank, among others, has called this assumption into question. One of John Maynard Keynes' main points in the *The General Theory* (1936) was that the existing levels of inequality were a primary cause of the depression and barrier economic recovery. "The outstanding faults

of the economic society in which we live” Keynes wrote, “are its failure to provide for full employment and its arbitrary and inequitable distribution of wealth and incomes” (Keynes 1936:372). He argued that the two were connected, that the high levels of wealth inequality contributed to the high levels of unemployment, mostly as a result of necessity of keeping interest rates high so as to ensure a high value of capital assets (by artificially keeping them scarce). Furthermore, the high rate of savings of the wealthy created a drag on aggregate demand. In fact, he states that “the growth of wealth, so far from being dependent on the abstinence of the rich, as is commonly supposed, is more likely to be impeded by it. One of the chief social justifications of great inequality of wealth is, therefore, removed” (Ibid.:373). High rates of inequality generate greater macro-economic instability, as they require higher rates of investment and other “injections” into the economy (government deficit spending) to make up for “leakage” from the affluent’s high savings rates.

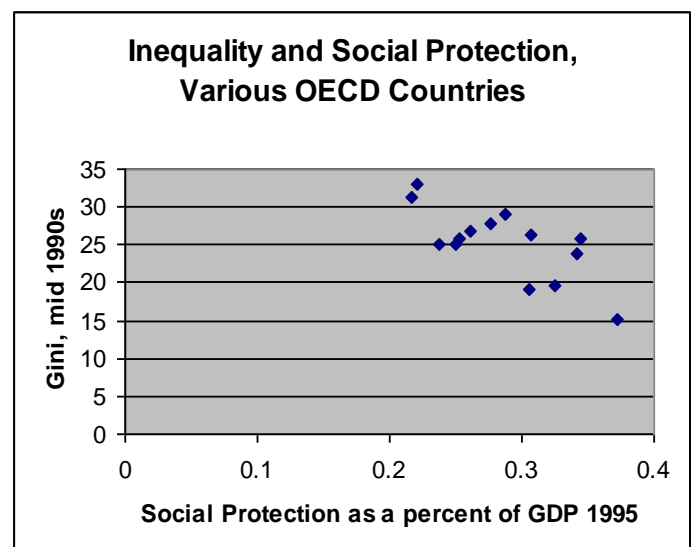
Governance Issues

The state has always been an important factor in determining the level of inequality, in fact, it was Adam Smith who stated: “Civil government, ... is in reality instituted for the defence of the rich against the poor, or of those who have some property against those who have none at all” (1976b:715). Besides the role of protecting the inequality *status quo*, government policies can increase inequality as well as to promote greater equality. All wealth is a form of property and thus exists partly because of government protections. Furthermore, legislation such as “limited liability” and “corporate personhood” help to shift some of the risks (and thus costs) of doing business away from capital and on to workers, consumers and

small businesses, thus promoting the wealth accumulation of the affluent.

Most attention on the role of government policy on wealth and income inequality concentrates on the influence of tax and spending policies, as well as on labor legislation and social welfare provisions. One the one hand you have the various subsidies, direct and indirect, that go to large corporations and the wealthy, plus their preferential tax treatment (capital income is usually tax at a rate lower than labor income, especially when one accounts for tax deductions) that clearly are an important aspect of the wealthy becoming and remaining such. However, on the other hand, governments also enact policies designed to promote greater equality, such as minimum wages, worker protections, recognition of legal status of unions, social welfare systems, universal education, support for health care and housing. It is fairly clear that the international differences in levels of income inequality are mostly the result of government policies and culture and not market forces. Particularly important are the levels of social protection provided by the state and the extent and level of minimum wage norms.

Graph 3. Inequality & Social Protection. OECD



Source: Adapted from OECD, LIS.

The discussion on the role of the state in terms of wealth and income inequality has centered on the concept of the trade-off between equality and efficiency. Originally developed by Arthur Okun (1975) to argue for greater efforts by governments to promote equality, the equality/efficiency trade-off has become the final argument against equality promoting policies. In order for a competitive market to work efficiently, Okun argued, it must be able to send out the correct price signals. Incomes are prices and thus any effort to manipulate market generated incomes will lead to distorted market signals, and thus inefficiency. In trying to divide the economic pie more evenly, the argument goes; such policies will inadvertently cause the economy pie to shrink, leaving less for everyone. However, the real world is much more complicated than the simple perfect competition model that the equality/efficiency trade-off is based upon. Robert Kuttner (1987) has shown that many equality promoting policies (such as universal education) also promote economic growth. Furthermore, while there is clearly a link between social protection levels and income inequality, no such link has been found between social protection efforts and lower economic growth rates (Pressman 2005).

Another argument supporting the link between economic growth and inequality states that high inequality promotes savings and that savings promotes capital investment and thus with higher rates of capital investment will come higher growth rates. However, this argument ignores the way capital investment is financed in a modern economy. As the US economy has shown over the past three decades, households do not have to save in order for businesses to invest.

Globalization and the “new economy” have put inequality on the top of the

economic research agenda. Many have argued that increased inequality is a necessary aspect of globalization (Wood 1996). Globalization has produced greater capital mobility, which has given capital greater market power, thus leading to higher factor incomes for owners of capital. Furthermore, the “new economy” requires greater labor flexibility, which typically has been translated into less protection for workers, decreasing their market power, and thus their factor incomes. Given the standard “welfare state model” will it be possible, given current policies, to promote both equity and efficiency. A basic income policy, which guarantees all citizens an income as a right of citizenship, has been promoted in many European countries because it would allow for greater labor flexibility without sacrificing economic security, and it allows for a way to insure that the benefits of the “new economy” are more equitably shared (Standing 1999; Clark 2002). Over two hundred years of capitalist development have shown that societies must continually adapt their means of promoting greater equality, which is necessary for ensuring stable democracies, with the continually evolving economy.

Selected References

- Atkinson, A.B. (1983) *The Economics of Inequality*. Second Edition. Clarendon Press, Oxford.
- Blau, F. and Kahn, L. (1998) “International Divergences in Male Wage Inequality: Institutions versus Market Forces”, *Journal of Political Economy*, 104, 41, 791-837.
- Buchanan, James M.; Robert D. Tollison and Gordon Tullock. (1980) *Toward a Theory of the Rent-Seeking Society*. College Station, Texas: Texas A & M University Press.
- Clark, Charles M. A. (1996) “Inequality in the 1980’s: An Institutionalist

- Perspective", in William Dugger (Editor), *Inequality: Radical Institutional Perspective on Race, Gender, Class and Nation*. Westport, CT.: Greenwood Press, 197-222.
- Clark, Charles M. A. (2002) *The Basic Income Guarantee: Ensuring Progress and Prosperity in the 21st Century*. Dublin, Liffey Press.
- Clark, Charles M. A. (2005) "Wealth as Abundance and Scarcity", in M. Naughton; H. Alford; C.M.A. Clark and S. Cortright (Editors), *Rediscovering Abundance: Interdisciplinary Essays on Wealth, Income and their Distribution in the Catholic Social Tradition*. South Bend, Indiana: University of Notre Dame Press.
- Galbraith, James K. (1998) *Created Unequal*, New York: The Free Press.
- Hicks, John. (1939) *Value and Capital*. Oxford, Oxford University Press.
- Hicks, John. (1942) "Maintaining Capital Intact: A Further Suggestion", *Economica*, IX, 174-9.
- Kaldor, Nicholas. (1955) *An Expenditure Tax*. London, Allen and Unwin.
- Kaldor, Nicholas. (1956) "Alternative Theories of Distribution", *Review of Economic Studies*, 23, 2, 83-100.
- Kuttner, Robert. (1987) *The Economic Illusions: False Choices between Prosperity and Social Justice*. Philadelphia: University of Pennsylvania Press.
- Kuznets, Simon. (1955) "Economic Growth and Income Inequality", *American Economic Review*, 45, 1, March, 1-28.
- Medema, Steven G. (1991) "Another Look at the Problem of Rent Seeking", *Journal of Economic Issues*, 25, 4, December, 1049-65.
- Milanovic, Branko. (2002) "True World Income Distribution, 1988 and 1993", *Economic Journal*, January, 51-92.
- Mill, John Stuart. (1987) *Principles of Political Economy*. Edited and with an Introduction by Sir William Ashley. Fairfield, NJ: Augustus M. Kelley.
- Okun, Arthur. (1975) *The Big Trade Off*. Washington, D.C.: Brookings Institution.
- Pareto, V. (1896) "La courbe de la repartition de la richesse", in *Recueil publie par la Faculte Droit a l'occasion de l'exposition nationale Suisse*, Lausanne: Universite de Lausanne.
- Paxton, Will. (2002) *Wealth Distribution*. Washington DC: Institute for Public Policy.
- Phillips, Kevin. (2002) *Wealth and Democracy*. New York: Broadway Books.
- Pressman, Steve, (2005) "Income Guarantees and the Equity-Efficiency Trade-Off", *Journal of Socio-Economics*, February, 1, 83-100.
- Smith, Adam. (1976a) *The Theory of Moral Sentiments*. Oxford: Oxford University Press.
- Smith, Adam. (1976b) *An Enquiry into the Nature and Causes of the Wealth of Nations*. Oxford: Oxford University Press.
- Standing, Guy. (1999) *Global Labour Flexibility: Seeking Distributive Justice*. London: Macmillan Press.
- Thurow, Lester G. (1975) *Generating Inequality*. New York: Basic Books.
- Thurow, Lester G. (1999) *Building Wealth: The New Rules for Individuals, Companies, and Nations in a Knowledge-Based Economy*. New York: Harper Business.
- Veblen, Thorstein. (1899) *The Theory of the Leisure Class*. New York, Macmillan.
- Veblen, Thorstein. (1904) *The Theory of Business Enterprise*. New York: Kelley & Co, 1965.
- Veblen, Thorstein. (1908) "On the Nature of Capital I and II" reprinted in Thorstein Veblen, *The Place of Science in Modern*

- Civilization and Other Essays*, 1990, New Brunswick: Transaction Publishers.
- Wilkinson, Richard G. (1996) *Unhealthy Societies: The Afflictions of Inequality*. London: Routledge.
- Wolff, Edward N. (1997) *Economics of Poverty Inequality and Distribution*. Cincinnati: South-Western College Publishing.
- Wolff, Edward N. (2002) *Top Heavy*. New York: New Press.
- Wood, Adrian. (1996) *North-South Trade, Employment and Inequality*. Oxford: Clarendon Press.

Websites

- University of Texas Inequality Project:
utip.gov.utexas.edu
- United Nations Development Programme.
www.undp.org
- Inequality.org. www.inequality.org
- World Bank, Poverty Network.
www.worldbank.org/poverty/inequal/data.htm

Charles M.A. Clark
Peter J. Tobin College of Business
St. John's University, New York, USA
New York City, USA.
clarkc@stjohns.edu

Informal Economy

Alys Willman-Navarro

Introduction

Common theories of economic development make little room for street hawkers, rickshaw pullers, and gypsy cabbies. Even less visible are the day laborers, informal workers in small factories and workshops or seamstresses who take in piecework from small textile factories. Their activities do not appear on official accounting sheets or factor into national income, and they are rarely addressed directly in policymaking. Yet these informal undertakings are stubbornly present everywhere, even in the most economically advanced environments.

Simply defined, the informal economy encompasses those economic exchanges that take place outside the sphere of formal regulation, where similar activities are regulated. Over time, the term *informal economy* has come to replace *informal sector*, as it was originally termed, because it better describes the myriad of informal workers and enterprises in both rural and urban sectors and their linkages to formal economic frameworks (see Chen 2004 for a detailed comparison of the two concepts).

The informal economy is often called the “underground”, “off-the-books”, “clandestine”, “invisible”, “unrecorded” or “black” economy. However these terms generally describe umbrella concepts for at least three kinds of economic activity: tax evasion on legally derived income; criminal activities (the illegal production and/or distribution of illegal commodities) and the informal economy. Most analyses exclude illegal economies from the informal sector, although others have argued that since informal transactions are unregulated they are in effect illegal. Castells and Portes (1989) have addressed this by focusing on the status

of the final product: that is, informal activities that produce a legal commodity are informal, whereas those that result in an illegal good are illegal - in effect, the qualitative difference between selling unlicensed childcare vs. selling crack cocaine. Estimates of the informal economy also generally exclude unpaid household production (domestic work and care activities) on the basis that although these activities have a market value, they are not produced directly for the market.

By definition the informal economy is difficult to measure, but researchers have estimated its size as equivalent to 41 percent of GDP in developing countries, 38 percent in transition countries and 18 percent in OECD countries in 2000 (Schneider 2002). In developing countries the informal economy continues to provide the bulk of employment opportunities, accounting for one-half to three quarters of all non-agricultural employment. In Sub-Saharan Africa the informal sector accounts for 78 percent of all non-agricultural employment (ILO 2002). In developed countries, off-the-books or unregulated work is increasingly common within the broader processes of deindustrialization and expansion of the services industry (Bernhardt & McGrath 2005).

It is the existence of a formal regulatory system that defines the characteristics of the informal sector – that is, the limits of the formal governance framework define the boundaries of the informal. This aspect is particularly important to those who view the informal sector as a refuge for entrepreneurs facing excessive state regulation (DeSoto 1989) or seek to measure its size, generally by focusing on non-regulated cash exchanges (Gutman 1977, Schneider & Enste 2000, Schneider 2002).

Because the size and nature of the informal economy depends directly on the particular institutional framework of a given country,

activities considered informal will vary from context to context. It is the existence of policies to govern economic activities that marks informal exchanges as part of a distinct historical process. That is, while informal “sweatshops” or unregulated childcare services may always have existed, the fact that health and labor regulations are now in place to regulate them differentiates them from those of 100 years ago (Sassen 1998). From this view, the informal economy is more than a set of activities. Rather, it is a *process* with changing boundaries, for which policy interventions will vary depending dependent on social context.

History of Informal Economy

Traditional theories of economic development considered the informal sector a residual, or temporary phenomenon that would disappear with industrial progress. Early modernization theorists predicted that the informal economy would gradually be incorporated into the modern industrial sector as development progressed (Lewis 1955). The “accelerated growth” models promoted within these viewpoints assumed that large scale industrialization would attract investment capital and pull workers from unproductive (informal, mostly agricultural) sectors of the economy toward the urban industrial (formal) sector, which would then generate resources to develop the larger economy and reduce poverty (Moser 1978). This belief was based on the experience of rebuilding Europe and Japan following World War II, and the expansion of industrialization in the United States and Britain.

Similarly, Marxist and Neo-Marxist theorists have located informal economic activity within the broader category of petty commodity production, which they predicted would gradually be eclipsed by the expanding capitalist sector. Capitalist production “destroys all forms of commodity production

which are based either on the self-employment of producers, or merely on the sale of excess product as commodities...” and, “by degrees, transforms all commodity production into capitalist production” (Marx 1972:36). Workers would thus be converted from commodity (subsistence) producers to formal wage laborers, a process described as “proletarianization.”

In the 1970s, the persistence and growth of the informal sector around the world began to call these predictions of the death of the informal sector into question. The International Labour Organization (ILO) was instrumental in providing empirical work to show that the formal economy was doing a poor job of absorbing surplus labor - a phenomenon economists began to call “jobless growth.” Rural to urban migration levels remained in excess of urban demand, suggesting that alternate sources of employment were drawing people from the countryside.

The original term “informal sector” was coined by anthropologist Keith Hart in a study of urban labor markets in Accra, Ghana (Hart 1973). Noting the contradiction between commonly-held Western theories of economic development and the dynamics he witnessed in Accra, Hart questioned whether the ‘unemployed’ were really a passive, exploited reserve army, or if they could be considered agents of growth. Hart presented a dualist model of income-earning activities based on a distinction between wage earning and self-employment, applying the concept of informality to the second category. In Hart’s view, the informal sector constituted an autonomous economic sector generating its own demand and supply of goods and services, and thus could be considered a source of potential growth. Hart recommended that governments capitalize on the dynamism and entrepreneurialism of the informal sector by making more efficient use

of manpower, preferably by implementing policies to pull informal workers into formal jobs.

The two-sector model proposed by Hart became known as the *Dualist* perspective, and was quickly taken up by development institutions, namely the World Bank and the ILO. Unfortunately, while the dual sector economy concept was adopted, Hart's vision of the informal sector as a dynamic force for growth was largely lost within the development bureaucracy. Instead, the growing literature emphasized the barriers faced by informal entrepreneurs. The informal sector was re-conceptualized as de-linked from the formal and limited by backward technology and small-scale operations, family ownership of firms, low-skill jobs, labor-intensive production and unregulated and competitive markets (Peattie 1980). Some have charged that the informal sector thus became considered synonymous with poverty and underdevelopment (see Castells & Portes 1989; Portes 1994:426-7, for critiques). In this context, studies of informal economies in developing countries proliferated to such a degree that Moser called them a "growth industry", while very few researchers studied this phenomenon in advanced economies (exceptions include Sassen's work on New York City (1989), Gershuny 1978, Pahl 1980).

Current Debates

Hart's view of the informal sector as a dynamic and autonomous component of the economy resurfaced in 1989 in a study by Peruvian businessman Hernando DeSoto on Lima's informal economy. DeSoto presented the informal sector as the "true market" driven by a vibrant group of, mostly rural-urban migrant entrepreneurs who, excluded from the formal employment and legal system of governance, set up their own businesses and institutions. His perspective is known as

the *Legalist* approach to the informal economy. In his book, *The Other Path* (1989) DeSoto challenged the negative conceptualization of informal workers, common in development policy circles, by describing them as rational decision-makers who weigh the costs of compliance with excessive regulations against the benefits of remaining outside the regulatory sphere. DeSoto highlighted the ways informal workers form parallel institutions that enforce contracts, create jobs and mediate disputes. Rather than a safety net for surplus labor, DeSoto saw the informal sector as a refuge for spirited entrepreneurs from the barriers presented by a bloated, inefficient bureaucracy. His recommendation was to promote informal entrepreneurialism by removing government regulations.

DeSoto's legalist view is very much in-line with a pure neoclassical view of the informal economy as adding dynamism to energize the formal economy. By offering goods and services not provided by the formal system, the informal economy is seen as providing competition and contributing to efficiency, job creation and growth.

Other Legalists are more skeptical, seeing the existence of the informal economy as a direct challenge to the welfare state's capacity to govern the economy (Schneider & Enste 2000). The expansion of the informal economy contributes to a "vicious cycle" whereby heavy tax burdens and regulations create incentives for firms to operate informally, which increases the size of the informal economy and reduces tax collection. This in turn increases the strain on public financing, resulting in even higher taxes and more incentive to avoid them. The end result can be a loss of confidence in public institutions, which undermines their ability to govern. Some have also posited that the vicious cycle generates a reduction in overall growth, because informal activities utilize

public infrastructure without contributing to tax revenue (Loayza 1996).

In contrast, *Structuralists* reject the dual sector view in favor of a structural explanation, considering the informal economy as an interdependent set of production relationships with the common feature that they serve to maintain the dominance of the capitalist sector. This view, rooted in Marx's concept of "petty commodity production" (Moser 1978) posited that rather than a formal/informal dichotomy, economic activities exist along a continuum of production processes defined by their relationships to the capitalist sector. By employing a vulnerable, low-skill labor force at low wages, the informal economy indirectly subsidizes the capitalist sector. The persistence and growth of the informal economy is thus more a result of the nature of capitalist development than a lack of overall growth. Moreover, estimates of the size of the informal economy in some regions was shown to remain constant during economic growth, and expand during crises, further reinforcing the claim that informality was a permanent feature of capitalist economies (Portes & Sassen 1987).

In conceptualizing the informal economy as a process with changing boundaries, structuralists aim to shift the focus from a set of activities toward the processes and dynamics that make them possible (Portes et al 1989). These analyses have sought to locate the informal economy within the broader process of de-industrialization in the post-war period (Sassen 1998, Portes & Sassen 1987). Especially during the 1990s, economic globalization brought incentives for companies to informalize their workforce in many industries and countries. In developing countries, this process formed part of the traditional manner of labor allocation in the economy. Benería (1989) for example, has shown how Mexican factories opted to keep

production costs low by further capitalizing on already flexible production arrangements. Instead of investing in technology or other options to increase productivity, these firms used lower labor costs from an informal workforce as a comparative advantage in the international market.

Feminist Debates

Given that women constitute the majority of informal workers everywhere, the informal economy has posed important questions for feminists. Indeed, research into the informal economy went a long way toward raising awareness about women's economic contributions more generally, particularly in developing countries (Moser 1977, Benería & Roldan 1987). Reflecting the larger policy debates, feminists have generally divided themselves into the Dualist and Structuralist camps (Bernasek 1999).

Dualist feminists have regarded the informal economy as a potential means of empowerment for women. By obtaining an income from informal work, it is assumed that women can increase their household bargaining power and better ensure the welfare of their children. The informal economy is often better able to accommodate women's multiple roles as household managers and community members by offering flexible schedules (see for example Horn 1994). In this way, informal activities are an important safety net for women excluded from formal jobs. The dualist feminist view has been promoted by international institutions including the United Nations, World Bank and USAID through projects including micro-credit, micro-insurance, training and technical assistance.

Structuralist feminists, in contrast, have focused on the vulnerability of women in the informal economy (Kalpagam 1994). Policies to empower women in the informal sector, it is argued, will only keep them there, where

capital is scarce and technology limited. Further, the structuralist feminist perspective regards women's informal employment as a means of subsidizing male wage labor in the formal economy by allowing firms to pay wages too low to support a family. Rather than supporting informal employment directly, structuralist feminists argue that governments should implement policies to create more high-quality, formal job opportunities for women.

One issue that has caught the attention of both dualist and structuralist feminists is the role of organizing women in the informal economy. Whereas traditional methods of organizing by formal labor unions have met with little success, independent organizations like the Self Employed Women's Association (SEWA) in India and the international organization Women in the Informal Economy Globalizing and Organizing (WIEGO), work to support women in the informal economy while also advocating structural change. By combining direct services such as micro-credit, micro-insurance and marketing assistance with advocacy and statistics collection, these organizations have the potential to address concerns from both feminist perspectives.

Issues in Measurement

Developing appropriate policies to address the informal economy and workforce depends on designing innovative approaches to measuring its size and contribution, and understanding the workers employed by it. A clear understanding of the contribution of informal activities to GDP is essential to improving models for national economic performance. Similarly, understanding the informal labor force is the key to designing adequate labor market policies (ILO 2002). In short, measuring and understanding the informal economy is essential to

understanding the full economy and developing appropriate policies.

Ethnographic work on the informal economy has generated a rich literature on its dimensions, functions and relationships in diverse contexts. But because of its very nature of operating outside formal mechanisms, measurement of the informal economy has been problematic. Classifications differ across countries, as does the quality and consistency of statistics collection, greatly hindering the capacity for international comparisons. There is no uniform method of measurement across different contexts. Several methods of measurement are generally employed, which can be classified as enterprise-based and activities-based measurements.

Enterprise-based measurements are often used by economists and macro-policymakers interested in quantifying the economic transactions taking place outside the regulatory framework by focusing on the firms that operate informally. These include, first, direct methods through surveys or tax auditing. While surveys provide useful information about consumption, they often rely on self-reporting, leaving them vulnerable to inconsistencies. In addition, surveys do not capture informal production or irregular labor in formal enterprises. Tax auditing methods make estimates based on the discrepancies between reported income and tax audit reports, but do not allow for random sampling. Thus, these direct methods provide information on the structure of the informal economy, but do not allow for reliable estimates of its size and growth.

Indirect, enterprise-based methods include macro-economic discrepancy methods, and electricity consumption methods. Discrepancy methods use a proxy for total economic production, both formal and informal. Official GNP is then subtracted from this estimate of "real" GNP to arrive at

an estimate of the unofficial economy. The most common proxy is currency demand, a method initiated by Cagan (1958) and developed by Tanzi (1980) and Schneider and Enste (2000). Using econometrics, the total demand for currency is estimated over time and compared to the actual amount of money in circulation. The drawback to this method is its assumption that informal transactions take place in cash. The electricity consumption method uses electricity use as a proxy for overall economic activity, and compares it to official GDP. Problems with this method are that it leaves out activities that don't use a lot of electricity, technology has gotten more efficient over time and uses less electricity, and elasticity of electricity use is different across countries. (For a full discussion of the different measurement methods see Schneider & Enste 2000).

A third common indirect method is the Very Small Enterprise method, commonly used in the United States. This method is based on the assumption that informal activities mostly occur in small enterprises (less than 10 workers). The VSE method has been criticized for two main reasons: first, it may overestimate the number of informal firms because not all VSEs engage in informal practices, and second, it may underestimate because fully informal VSEs escape government record-keeping (see Portes 1994 for discussion).

Recognizing the limitations of using only enterprise-based approaches, the ILO has called for further work on researching employment or activity-based estimates of informality. Activities-based approaches rely on census data and labor market surveys to compare the official labor force with estimates of the total labor force (usually extracted from the number of self-employed workers, who constitute the major component of informal employment). Activity-based methods are assumed to give a more accurate

picture of the diversity of activities and workers in the informal economy by using the individual as the point of departure. The obvious drawback to these is that they assign workers to either the formal or informal labor markets, which incorrectly reflects the activities of workers who participate in both economies.

Estimates of Size and Contribution

Experiences in the last two decades have led to recognition of the informal economy as a much more diverse and dynamic set of activities than previously thought. There exists now some convergence of views that the informal economy is a continuum of economic relations (from formal and regulated to informal and unregulated) and that it is highly segmented into different types of activities. The latter range from employers to self-employed operators, unpaid family workers, employees of informal firms and industrial homeworkers (see Chen 2004 for analysis). Together these activities represent a significant contribution to total economic output – from 13 percent in terms of GDP in developed countries to 42 percent in Africa, on average.

*Table 1: Informal Economy
Percent of GDP, 1999/2000*

Africa	42
Asia	26
Latin America	41
Europe (OECD)**	18
North America	13
Australia & NZ	14

Source: Adapted from Schneider 2002

*Unweighted averages. **Estimate based on 16 European OECD member countries.

In developing, transition and developed countries the informal economy provides an important source of jobs, goods and services. As shown in the table below, the informal economy accounts for roughly one half to

three quarters of all non-agricultural employment in developing regions. In Sub-Saharan Africa 78 percent of the total workforce and 91 percent of the female labor force are informally employed. In the Middle East – Northern Africa region, these numbers tend to be comparatively lower because relatively high shares of the female workforce are employed as formal employees in the public sector and as unpaid contributing family workers.

Table 2. Informal Non--Agricultural Employment. Percent of Total Non-Agricultural Employment, 1994-2000

Region	Total Workforce	Female	Male
Latin America	51	58	48
Sub-Saharan Africa	78	91	72
MENA	47	42	49
S & SE Asia	71	73	72

Source: Adapted from ILO (2002)

Notes: (1.) Non-weighted averages for 10 countries in Latin America; 4 countries in Sub-Saharan Africa (excluding South Africa); 1 country in the Middle East plus 4 in North Africa; and 1 country in South Asia plus 3 in Southeast Asia. (2) In the MENA region, relatively high shares of the female workforce are a) formal employees in the public sector and b) unpaid contributing family workers.

Women make up the majority of informal workers everywhere. Around the world, women account for 30-90 percent of street vendors, 35-80 percent of all home-based workers and 80 percent of domestic workers (ILO 2002). Women are also significantly over-represented in the number of self-employed workers (the majority of whom operate informally) representing 72 percent in North Africa, 71 percent in Sub-Saharan Africa, 60 percent in Latin America and 59 percent in Asia (ILO 2002). Women tend to occupy the low-skill, low-wage segments of the informal economy, especially as domestic or unpaid family workers, resulting in a

significant gender gap in wages (see Chen et al 2004 for detailed gender analysis).

The disproportionate number of women employed informally stems from several global demographic and economic trends. These include, first, an overall increase in women's labor force participation across nearly all regions of the world; increased urbanization and changing gender norms. Broader economic trends stemming from shifting global production systems include increased demand for female labor in certain industries, especially small-scale industry and homework; rising demand for low-skill labor generally; and persistently high unemployment and underemployment, especially in developing regions.

In developed countries, women make up a disproportionate number of part-time or irregularly-employed workers. Women in OECD countries represent 60% or more of part-time workers. The percentage of women employed part time is 98% in Sweden, 80% in UK, and 68% in Japan and the United States (ILO 2002).

Table 3. Women in Informal Economy, 2002

Region	Percent of Women in Informal Employment
Latin America	58 (48% of men)
Sub-Saharan Africa	84 (63% of men)
North Africa	43
Asia	n/a
Worldwide	60

Source: Adapted from ILO (2002)

Global Trends

The fact that a significant portion of economic activity escapes formal regulation everywhere is a matter of global importance and presents clear challenges to governance. As the persistence and (in many cases) the growth of the informal economy has been recognized and documented, researchers have sought to locate these trends within the broader context of economic globalization.

Chen (2004) has identified three types of economic growth that contribute to informalization: jobless growth, in which the economy does not create enough jobs to meet demand, pushing workers to seek supplemental sources of income; high-tech growth, in which new jobs are located in high-skill sectors of the labor market, leaving lower skill workers with limited formal opportunities; and “growth from below”, in which new jobs are created primarily in the small and microbusiness sectors, many of them in informal enterprises.

Changes in global production structures also drive informalization, as firms segment production across different geographical locations. Often, developing countries and regions use informalization as a comparative advantage in attracting foreign capital. They may specialize in particular parts of the process, where firms often look for more flexible arrangements and/or more lax environmental and health regulations. This gives rise to more subcontracting and home-based work arrangements.

Studies of the informal economy in both developing and industrialized contexts have highlighted the *anti-cyclical* nature of informal activities - that is, parts of the informal economy tend to grow with a decline in aggregate demand, and decline with its growth. In contrast to the formal employment sector where reduced aggregate demand tends to shrink formal employment, employment in the informal economy is largely supply-driven, so that new entrants are relatively easily absorbed into an expanding informal economy.

As examples, the experiences of Latin America in the 1980s (Tokman 1992) and Asia in the 1990s (Lee 1998) highlight the role of the informal economy as a buffer against the impacts of crisis and/or reform. In both regions, the informal economy expanded significantly during the economic crises,

providing an important safety net for workers. Several common factors tend to drive this trend. First, rising inflation and currency crises drive up the cost of living, pushing workers to supplement or replace real lost income. Second, firms facing rising competition and production costs seek to downsize and/or informalize their workforce through more flexible, often irregular, employment arrangements.

Likewise, economic structural reform in poor or transitional economies can feed the expansion of informal employment through downsizing of the public sector and cuts in basic public services, which raise the cost of living. Some have pointed to IMF austerity programs as contributing to informality by raising taxes, providing incentives for firms to exit the formal economy.

In developed countries as well, the informal economy has remained constant and even expanded during economic downturns. However, in contrast to developing countries, where large informal economies are part of a historical manner of allocating labor and capital, in developed countries the proliferation of informal production arrangements represents a return to past practices (Portes and Sassen 1987). Several explanations have been offered. First, the encroachment of union power has been signaled as a disincentive for firms to operate formally, as in the case of many northern Italian firms during the 1970s (Brusco 1982, Capecchi 1989). However, the opposite has also been true, as unions serve to keep firms from shifting production to the informal economy in highly unionized industries and contexts, such as the US automobile industry. Another partial explanation points to increased competition from developing country exports and the presence of a low-cost, mostly immigrant pool of labor to explain firms' decisions to informalize production in wealthier countries (Ybarra

1982). The latter does not sufficiently explain why informal arrangements have expanded in countries where migration levels are comparatively low, as in Spain and Italy during the 1990s.

Increasingly, research on the informal economy in developed countries explores not only industrial trends toward subcontracting and other flexible arrangements, but also the growing informalization within the expanding services sector more generally (Bernhardt and McGrath 2005). This helps explain the proliferation of unregulated work in nail salons, furniture and car workshops and other industries that have never been highly unionized. Other work has highlighted the role of informal production in rural areas where manufacturing plants have closed or moved elsewhere, as in the U.S. Appalachian region (Oberhauser 2002). One particular growth sector, especially in the United States, has been the expansion of the home-improvement and small-scale construction industry that employs a disproportionate number of casual or day laborers (Valenzuela and Meléndez 2003).

Policy Responses and Governance Issues

The concept of the informal economy continues to attract the interest of policymakers and researchers who seek to understand the myriad of activities it represents. As the size and significance of the informal economy not simply as a refuge from poverty but as a provider of jobs, goods and services, has been recognized, several key governance issues have arisen. As a cross-cutting policy issue, discussions of policy toward the informal economy have been incorporated into theories of civil society and institutions as well as more mainstream theories of economic development more generally.

As Chen (2004) has noted, current debates on the informal sector tend to divide into two

foci: those that focus either on informal employment that is not *protected* (leaving workers vulnerable to low wages and inadequate working conditions) or informal firms that are not *regulated* (and associated with unfair competitive advantages through their avoidance of taxes and legal regulations). With the Dualist approach now considered outdated, the current debate on governance falls largely along Legalist vs. Structuralist lines.

A central question here is the practical meaning of "formalizing", informal activities and/or workers. Policymakers, concerned with bringing informal activities within the taxable, regulatory system, often argue that informal firms should be required to comply with formal licensing, registration and tax regulations. Legalists, departing from the experience of micro-entrepreneurs and small businesses, argue for the removal of some of these barriers, such as registration, in order to give informal entrepreneurs access to the benefits of the formal system, protection of contracts, legal title to land and other capital.

Structuralists focus on extending the social protections of the formal governance system to currently unregulated work. They call for a new "social contract" between the state and society that breaks the links between jobs and social benefits. This implies not only enforcing accountability of firms to workers, but also the provision of more comprehensive social protection by the government in order to reduce worker dependence on employers. It also means enlisting government support in creating an enabling environment for business and creation of more formal, protected jobs (Portes et al 1989).

These competing recommendations highlight an important policy challenge best articulated by the ILO (2002b) as the "dilemma of the informal sector." That is, should governing bodies seek to promote the informal sector as a provider of employment

and incomes, or try to extend regulation and protection to it, which then implies possibly reducing its capacity to provide jobs and income for the workforce?

As a response, Chen has proposed an alternative approach that seeks to apply the analysis of the three main schools of thought to different segments of the informal economy where they may be more effective. That is, some households have few links with the formal regulatory environment (dualist) while other workers are involved in activities that place them in a subordinate position to larger firms (structuralist). Other informal entrepreneurs look to escape regulation as a means of growing their enterprises (legalist) and would thus benefit from relaxing legal restrictions. With this in mind, Chen has argued for designing appropriate responses based on the needs of each segment of the informal economy, rather than applying a single policy response to all sectors (Chen 2004). By recognizing the diversity of experiences within the informal economy, a more comprehensive approach may better address the needs of more vulnerable groups, especially female workers.

The role of non-governmental actors is increasingly important in developing policy approaches. Activists and organizers of informal workers play a key role in raising awareness about informality, advocating for more effective policies, and addressing key needs such as linking informal workers to international markets. Over the years these groups have filled gaps for services left by the state while advocating for structural change. Their role is often seen as helping to bridge the divide between the Dualist/Legalist approaches and Structuralist approach. Feminists, in particular, have embraced this middle ground approach as a viable way forward for women in the informal economy.

Organizations like India's Self Employed Women's Association (SEWA) formed

originally to support immediate needs of informal, mostly female, workers. Over time SEWA has added a range of services, beginning with micro-credit and later expanding to micro-insurance and technical assistance. In 1997 SEWA joined with HomeNet, an international organization of home-based workers, as well as researchers and statisticians to form Women in the Informal Economy Globalizing and Organizing. Among other projects, WIEGO works to connect informal entrepreneurs with other workers around the world using the Internet. In addition, WIEGO links informal producers with international markets for their goods. In the United States, efforts to organize collectives of home-based workers have also contributed to improved working conditions and empowerment (see Oberhauser 2002 for evidence from Appalachia).

Critical Issues for the Future

As recognition of the contribution and importance of the informal economy has expanded, interest in understanding its dynamics has grown, as have initiatives to support its activities. One of the most important challenges facing policymakers and researchers continues to be how to improve measurement of the informal economy in order to better inform policy initiatives and the work of non-governmental actors. Toward this end, the ILO has begun to promote increasing collaboration between statistics collectors and users (activists and organizers) to improve statistics and analysis for policymaking (ILO 2002).

A second concern, particularly for organizations of informal workers, is connecting informal entrepreneurs with national and international markets. Within the broader movements for fair trade and ethical trade, organizations of informal workers have developed projects to sell their products by

plugging into international networks and using Internet-based technologies. Strengthening these linkages, building capacity among informal workers, and raising awareness among consumers are key challenges moving forward.

For governments, addressing the challenges of growing informality means arriving at the appropriate mix of policy interventions to support informal workers while strengthening the formal institutional environment. These will vary depending on context and could range from better enforcement of labor protections and simplification or reform of tax systems, to broader changes in employment regulations or social protections.

Thus, the challenge for all institutions is to use creative approaches to better capitalize on the wealth of human and economic resources represented by the informal economy.

Selected References

- Benería, Lourdes and M. Roldan. (1987) *The Crossroads of Class and Gender: Homework, Subcontracting and Household Dynamics in Mexico City*. Chicago: University of Chicago Press.
- Benería, Lourdes. (1989) "Subcontracting and Employment Dynamics in Mexico City", in A. Portes; M. Castells and L. Benton. (Editors), *The Informal Economy: Studies in Advanced and Less Developed Countries*. Baltimore, MD: The Johns Hopkins University Press.
- Bernasek, Alexandra. (1999) "Informal Sector", in Janice Peterson and Margaret Lewis (Editors), *The Elgar Companion to Feminist Economics*. Cheltenham, UK and Northampton, MA: Edward Elgar.
- Bernhardt, Annette and Siobhan McGrath. (2005) *Trends in Wage and Hour Enforcement by the U.S. Department of Labor, 1975–2004*. New York: Brennan Center for Justice.
- Brusco, S. (1982) "The 'Emilian' Model: Productive Decentralization and Social Integration", *Cambridge Journal of Economics*. Vol 6, 167-84.
- Cagan, Phillip. (1958) "The Demand for Currency Relative to the Total Money Supply", *Journal of Political Economy*, 66, 4, 303-28.
- Capecchi, V. (1989) "The Informal Economy and the Development of Flexible Specialization in Emilia-Romagna", in A. Portes, M. Castells, and L.A. Benton (eds.), *The Informal Economy: Studies in Advanced and Less Developed Countries*, Baltimore, MD: The Johns Hopkins University Press.
- Castells, Manuel and Alejandro Portes. (1989) "World Underneath: The Origins, Dynamics and Effects of the Informal Economy", in A. Portes; M. Castells and L. Benton (Editors), *The Informal Economy: Studies in Advanced and Less Developed Countries*. Baltimore, MD: The Johns Hopkins University Press.
- Chen, Martha. (2004) *Rethinking the Informal Economy: Linkages with the Formal Economy and the Formal Regulatory Environment*. Paper presented at the Expert Group on Development Issues and United Nations University World Institute for Development Economics Research conference, 17-18 September, Helsinki, Finland.
- Chen, Martha; J. Vanek and Marilyn Carr. (2004) *Mainstreaming Informal Employment and Gender in Poverty Reduction: A Handbook for Policy-Makers and Other Stakeholders*. London: Commonwealth Secretariat.
- DeSoto, Hernando (1989) *The Other Path: The Invisible Revolution in the Third World*. New York: Harper and Row.
- Gershuny, Jonathan (1978) *After Industrial Society: The Emerging Self-Service Economy*. London: MacMillan.

- Hart, Keith. (1973) "Informal Income Opportunities and Urban Employment in Ghana", *The Journal of Modern African Studies*, I, p. 61-89.
- Horn, Nancy E. (1994) *Cultivating Customers: Market Women in Harare, Zimbabwe*. Boulder and London: Lynne Rienner Publishers.
- International Labour Office. (1972) *Employment, Incomes and Equality: A Strategy for Increasing Productive Employment in Kenya*, Geneva, ILO.
- International Labour Office. (1993) *ILO Report of the Fiftieth International Conference of Labour Statisticians*. Geneva: ILO.
- International Labour Office. (2002) *Women and Men in the Informal Economy: A Statistical Picture*. Geneva: ILO.
- International Labour Office. (2002b) *Decent Work and the Informal Economy*. Report VI, Geneva.
- Kalpagam, U. (1994) *Labour and Gender: Survival in Urban India*. New Delhi: Sage Publications.
- Lee, E. (1998) *The Asian Financial Crisis: The Challenge for Social Policy*. Geneva: International Labour Organization.
- Lewis, W.A. (1955) *The Theory of Economic Growth*. London: Allen and Unwin.
- Loayza, Norman. (1996) "The Economics of the Informal Sector: A Simple Model and Some Empirical Evidence from Latin America", *Carnegie-Rochester Conference Series on Public Policy*, 45, 129-62.
- Marx, Karl. (1972) *Capital*, 2. London: Lawrence and Wishart.
- Moser, Caroline. (1977) "The Dual Economy and Marginality Debate and the Contribution of Micro Analysis: Market Sellers in Bogota", *Development and Change*, 8 2, 465-89.
- Moser, Caroline. (1978) "Informal Sector or Petty Commodity Production: Dualism or Dependence in Urban Development?" *World Development*, 6, 9/10, 1041-64.
- Oberhauser, Ann M. (2002) "Relocating Gender and Rural Economic Strategies", *Environment and Planning*, 34, 7, 1221-37.
- Pahl, Raymond. (1980) "Employment, Work and the Domestic Division of Labor", *International Journal of Urban and Regional Research*, Vol 4, March, 1-20.
- Peattie, Lisa. (1980) "Anthropological Perspectives on the Concepts of Dualism, the Informal Sector and Marginality in Developing Urban Economies." *International Regional Science Review*, 5, pp 1-31.
- Portes, Alejandro and Saskia Sassen. (1987) "Making it Underground: Comparative Material on the Informal Sector in Western Market Economies", *American Journal of Sociology*, 93, 1, 30-61.
- Portes, Alejandro; Manuel Castells and Lauren Benton. (1989) *The Informal Economy: Studies in Advanced and Less Developed Countries*. Baltimore and London: Johns Hopkins University Press.
- Portes, A. (1994) "The Informal Economy and Its Paradoxes", in Neil J. Smelser and Richard Swedberg (Editors), *Handbook of Economic Sociology*. New York: Russell Sage.
- Sassen, Saskia. (1989) "New York City's Informal Economy", in Alejandro Portes, Manuel Castells and Lauren Benton (Editors), *The Informal Economy: Studies in Advanced and Less Developed Countries*. Baltimore and London: The Johns Hopkins University Press.
- Sassen, Saskia. (1998) *Informalization in Advanced Market Economies*, Issues in Development Discussion Paper 20. Geneva, International Labour Office.
- Schneider, Friedrich and D.H. Enste. (2000) "Shadow Economies: Sizes, Causes and Consequences." *Journal of Economic Literature*. 38, 1, March, 77-114.

- Schneider, Friedrich. (2002) *Size and Measurement of the Informal Economy in 110 Countries Around the World*. Washington DC, World Bank Working Paper.
- Tanzi, Vito. (1980) "The Underground Economy in the United States. Estimates and Implications", *Banca Nazionale del Lavoro Quarterly Review*, 135, 428-53.
- Tokaman, V. (1992) (Editor) *Beyond Regulation: The Informal Economy in Latin America*. Boulder CO: Lynne Rienner Publishers.
- Valenzuela, Abel and Edwin Meléndez. (2003) *Day Labor in New York: Findings from the New York Day Labor Survey*. New York, April 11.
- Ybarra, Josep (1982) "Economía Subterránea: Reflexiones sobre la Crisis Económica en España", *Economía Industrial*, 218, 33-46.

Websites

- International Labor Office, Employment Sector Division, Informal Economy. www.ilo.org/public/english/employment/infec/index.htm
- Women in Informal Economy Globalizing and Organizing. www.wiego.org
- Self Employed Women's Association. www.sewa.org

Alys Willman-Navarro
World Bank
Washington DC.
USA
AWillman@worldbank.org

Information and Communications Technology

Wilfred Dolfsma and Ferdinand Jaspers

Introduction

Information and Communications Technology (ICT) is a set of technologies, combining hardware (computers, switches, mobile devices, etc.) and software (data processing applications, office solutions, operating systems, etc.), which enables agents to process data and communicate with each other. Agents can be private individuals, organisations as well as machines. Computing technology and the infrastructures that connects this hardware can be used to operate a variety of programs for both business, social as well as entertainment purposes. The realization that the hardware is valuable only for what it allows users to do has meant that the software that used to be provided free of charge when a computer was bought is now likely to be sold separately.

The software not required for the hardware to function or the connections to be established may be called 'content.' Content may be entertainment, software for communication purposes, or software be used for a number of different other purposes. The latter include producing texts, analyzing data, using and manipulating databases, designing products, etc. Combinations of these are also feasible. Given the widespread possibilities for applying ICTs it can best be seen as a general-purpose technology. Computing technology is increasingly used in products that had already been in existence to make them 'smarter'. Examples of this include refrigerators and automobiles. In addition, ICTs have also been used extensively to change and speed up many processes with which a range of goods and services are produced. As such, its use has implications for many branches of industry and corners of

society. ICT has been characterized as a general purpose technology because of its wide-ranging effects.

The potential nature of its effects have been compared with those of the (two) Industrial Revolutions (from 1780 and 1860) by economic historian Joel Mokyr (2002) for one. Carlotta Perez (2002) and Dosi et al. (1988) have claimed that ICTs are causing a new technological paradigm that will affect the economies around the world fundamentally. Certainly since the introduction of the first personal computer by IBM and the privatization of the Internet in 1994 ICTs have had a tremendous impact on economies and societies.

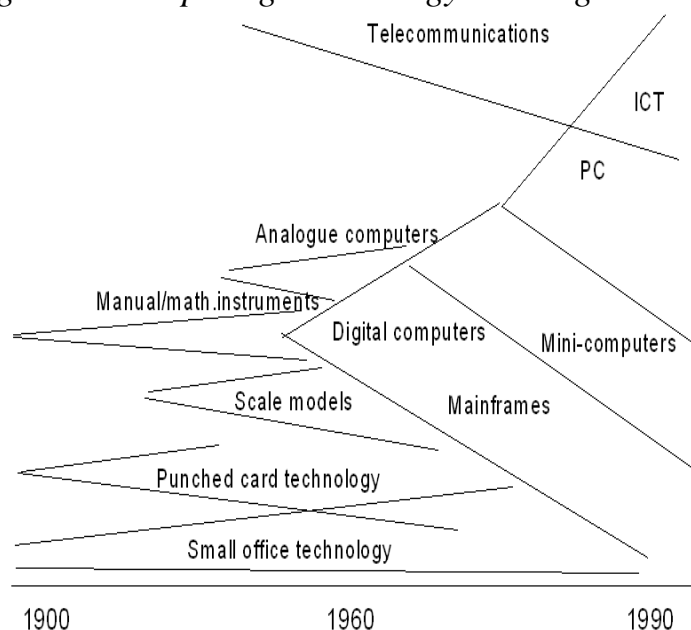
Development of ICT

Of course, one can trace the roots of computing technology even before the 1960s (Van den Ende & Dolfsma, 2005), and maybe even to Charles Babbage's Difference Engine of 1822. However, from the 1960s to date, ICTs have developed rapidly (see Figure 1; width of the angles roughly indicating relative dominance of the technology), mainly through autonomous developments in the technological knowledge needed. In the late 1960s the US defence department sponsored the construction of a computer network between several universities to allow for communication. ARPANET as it was called grew rapidly. In the early 1990s, the combination of computing technologies and communications technologies has given the paradigm a boost and has made it available to a much wider audience. The use of the computer began to spread throughout economy and society rapidly due to the privatization of the Internet in 1994 too.

At the end of 2003 the penetration rate of personal computers in households was over 50% in most developed countries, and the number of fixed Internet connections in OECD countries was about 259 million

(OECD 2005). In addition, both numbers are growing rapidly and also the quality of Internet connectivity is improving, since users increasingly access the Internet through 'always on', high-bandwidth connections (broadband Internet).

Figure 1. Computing Technology Paradigms



Source: Adapted from Van den Ende & Dolfma (2005) Note: Slopes indicate market growth

What is truly amazing is the prediction by co-founder of computer chip manufacturer Intel, Gordon Moore, that the number of transistors on a silicon chip would double every 18 months or so has held up to date. An important proviso for the economically minded is that 'Moore's Law' assumes that the price for a chip would not rise. Little known is Moore's so-called second law. This claims that the costs of development to realise the first law will rise so much that few firms will be able to invest the amounts of money needed. Oligopoly or even monopoly conditions may ultimately prevail.

The technological progress of ICTs is thus driven by an interplay of social, economic, political, and technological factors. Autonomous technological developments (technology push) constitute an important driver for advances in ICT, e.g. in the field of

computer chips. Besides the importance of technological development, some periods of change are largely driven by increased demand (market pull) for ICT in the society at large (Van den Ende & Dolfma 2005). Political decisions in the 1980s and 1990s towards the liberalization of national telecommunications markets and the privatization of their incumbents further stimulated economic activity and the incentives of new and established firms to innovate (Shy, 2001).

ICTs and Long Waves

A new technological paradigm is characterized by the fact that an entirely new industry or set of industries emerges (Freeman & Perez 1988). The effect of technology on society has been thought of to wax and wane, much like waves. The first one to note this has been a Russian statistician by the name of Kondratieff (Kleinknecht & Van der Panne 2008). The waves that carry his name are said to last 40-50 years each. Depending on how one defines waves in this perspective that places much emphasis on the role of technology in shaping economy and society, 5 different waves may be distinguished to date. The first one would be called the (first) Industrial Revolution, where mechanisation occurred. The wave affects some industries more than others; certain countries are the technical as well as economic leaders of the world. The second wave was driven by steam power and the railways. The third one relied on electrical and heavy engineering. The fourth was one of Fordist mass production from the 1930/1940s towards the 1990s. The fifth Kondratieff wave is thought to be due to ICTs, starting in the 1990s. The timing of this wave coincides with the merger of information technology on the one hand, and communication technology on the other. Previously, the two had developed rather separately.

Not only are new industries associated with the widespread use of Information & Communications Technology (ICT), but as indicated above it has also changed the workings of other industries in fundamental ways. In this respect ICT may be said to represent a new technological paradigm. Countries, similarly, may be affected in different ways by ICTs. Some of the advanced countries such as the US will certainly be able to maintain their positions, others may see their (relative) position deteriorate. Some developing countries, not bothered by vested interests and substantial investments in existing infrastructure which may not be useful under new circumstances, will be able to catch up seizing the technological window of opportunity. One may think of South Korea, Taiwan and India in particular.

Economic Impact of ICT

ICTs have enabled a number of developments in the economy. Despite the popular belief in the 1990s that ICTs would create a New Economy, technologies changed rapidly, but some economic laws did not (Shapiro & Varian 1999). Discussion on this is, however, by no means concluded (Courvisanos 2004). It is important to remember that mainstream economics certainly does not vindicate whatever development ICT has led to—indeed, it can be used to critically evaluate for instance the role in which consumers may be forced (Dolfsma 2006).

The ICT sector has become an important part of the world economy however, accounting for about 10% of GDP in OECD countries (OECD 2004b). The ICT sector is also a very dynamic part of the world economy as it is characterized by fierce competition, high investments in R&D, and strong growth. In addition, spillovers occurred to associated economic activities, such as the construction of networks,

maintenance of information systems, and systems integration. Nevertheless, there has been a long discussion about why the effects that everybody experiences do not show up in productivity statistics at the national level: “You can see the computer age everywhere but in the productivity statistics.” This is sometimes referred to as the Solow paradox, due to remarks Nobel Laureate of economics Robert Solow has made in 1987 in the New York Review of Books. This is partly due to the way in which national accounts are compiled and partly also due to the qualitative changes ICTs engender, such as new goods and services that can now be offered. In addition, due to the fast rate with which computing technology develops, the economic deterioration of computers is quick too: well before computers stop to function technically, they may become useless economically and thus need to be replaced. Depending on what one procedure one uses to compile national accounts, this may underestimate the contribution of computers to the economy.

Developments in ICT also have a significant impact on economic activity in general, increasing the efficiency and speed of operations and promoting product and process innovations, for which relevant and timely information is a key input. More specifically, ICT contributes to the increasing importance of services in the modern economy. Certain services, such as financial services, have largely moved on-line. In this way traditional business models and distribution channels are surpassed in settings as diverse as tourism and selling books. More in general, the Internet provides a means to organize and match supply and demand, which may stimulate entrepreneurship. Thousands of people for instance make a living out of dealing on the eBay website or playing online games such as Second Life.

On-line business processes and transactions are referred to as e-business or e-commerce. Most firms present themselves on the Internet and communicate their product offerings and prices, but especially smaller firms are slow to perform or facilitate additional activities on-line, such as ordering and selling. This is especially true for more complex applications, such as the integration of IT systems between buyers and suppliers in the supply chain (OECD, 2004b). With the increased availability of ICTs, the challenge shifts more and more from the availability of ICT to the exploitation of ICT opportunities, both within firms and between firms or between firms and consumers.

ICT - through the increased availability of broadband connections - also triggers the convergence of the traditionally separated industries of telecommunications, media (e.g. content, broadcasting, magazines) and technology. As a result, services like video on demand and IP TV (television over the Internet) become possible. Although this trend prompted technology and media firms to integrate (e.g. AOL and Time Warner), clear evidence that such an approach results in significant synergies is (still) lacking.

ICTs have thus changed the structure of numerous industries. This has started with the music industry. The business model of firms in the music industry was based on the exclusivity of commercially exploiting the products they produced that copyrights allowed for. The issue of applicability of copyrights on the Internet is not settled yet – indeed, one may expect that major legal and political battles will be fought over this. Even if a similar protection will become available as before, enforcement may not be possible. As bandwidth grew, the Movie and TV businesses have had to re-consider their business models as well. More recently, the telecom industry has seen many of its customers start to use their connection to the

Internet as a means of making phone calls. Be it mere phone calls facilitated by the Voice over Internet Protocol (VoIP), or in combination with images and text-messaging.

Information Goods

Developments mainly emanate from the fact that ICTs have increased the information-content of products. Information goods (e.g. software, music) differ from physical products in a number of important respects (Shapiro & Varian 1999; Shy 2001).

For instance, most information goods are costly to produce (i.e. they require large initial, sunk investments), but cheap to reproduce (i.e. scale economies are large). As the protection provided by intellectual property rights is imperfect third parties cannot always be prevented from copying an information good. Parties other than the creator or her agent may appropriate the rents of innovation. Although the argument is often made that people may not be motivated to develop new information goods under those circumstances, this is not obviously true (Dolfsma 2006b). There certainly does not seem to be a lack of content on the Internet.

As the marginal costs of producing additional copies of an information good approaches zero, many economists would argue that it is in the public interest to diffuse this good as broadly as possible. The use of information goods may not be limited to certain individuals or groups while such use does not limit the use of others, making a good a public good. Information goods may be under-produced for exactly that reason. There would not be an incentive for firms to invest in the creation and reproduction of such goods, and yet many firms do. Open Source Software – software which is freely available to any user who wishes to, and may be developed further by anyone provided the extension is freely available too – however, is at least partly produced because of the

contribution by (employees of) commercial undertakings. Using existing copyright law, any contribution to, e.g., the Linux Operating System cannot be commercially exploited but needs to be made available to anybody who wants to use it. Such firms may, of course, sell complementary goods and services to make a profit.

Copying or using information goods does not affect their quality. Ease of reproduction means that there may be a lot of information goods for sale on the market that may be second hand goods, but are in fact indistinguishable from new goods. The sign of the price of second hand goods for the quality of the first hand good is thereby negated. There is then an increasing value on a firm's or organisation's reputation as the number of information goods as well as the amount of information available explodes.

Another characteristic of information goods is the interdependence between software and hardware. Both complementary elements need to be compatible for ICT systems to perform as intended. Compatibility for instance allows the exchange of files and programs between computers. The need for compatibility between hardware and software, and between different pieces of hardware and software themselves, results in the importance of standards, such as DVD for recorded media, and the GSM (Global System Mobile) standard in Europe or the CDMA (Code-Division Multiple Access) standard in the US for mobile telephony. These allow producers of complementary products to innovate autonomously.

The adoption of ICT standards is often characterized by network effects or externalities - the value of the standard increases significantly as the number of users grows. Certain hardware or software configurations such as the IBM personal computer or the Microsoft Windows operating system for desktops function as

platforms for a number of complementary goods and services. A larger installed customer base increases the possibility to communicate or share software, and offers incentives to providers of complementary goods and services to improve the elements of the system. If the specificities of the interfaces between a dominant platform and complementary goods and services can be controlled, some players that might want to enter a market may be prevented from doing so. Dominant players such as Microsoft and Google are chided for this practice.

Social Impacts:

Consumer, Civilian, Individual

Developments in ICT have dramatically altered the lives of millions of people. In less than two decades, individuals in most developed countries now have a PC, an always-on Internet connection, as well as a cellular phone. In addition, ICT is included in most consumer electronics and most products are available for reasonable prices. These developments also confront people with applications like e-mail, e-commerce, weblogs (sites where people relate about their personal lives and express their opinions), instant messaging (IM), computer games, and text messaging on mobile phones.

Especially for youngsters these latter applications have a strong influence on their social lives, making contacts and communicating with each other using electronic means. New developments in technology, especially those technologies that have an entertainment aspect to them, may have affected young people more than people in other age group. One may compare with radio and television (Dolfsma 2004). Still, ICTs allow for de-centralized communication, and so poses additional challenges to those individuals (e.g., parents) or parties (such as governments) that may wish to control what information and

entertainment is available. In addition, ICT has changed the education received. Modern educational systems make extensive use of PCs, digital content, the Internet, and interactive software. Finally, ICT has also improved health services. For the analysis and treatment of patients extensive use is made of the latest developments in ICT, for instance in MRI (magnetic resonance imaging) scans. Data on people's health situation can be digitally stored in a central place, accessible by many different parties, potentially increasing efficiency and preventing repeat tests. Not everybody's interests may be served by using these technical possibilities, however. A patient in a hospital may not like his insurance company to have access to every piece of her personal information, nor may the physician be willing to share information as it may undermine her position of privilege. An individual's privacy might be violated; an expert's opinion might be questioned by outsiders who are not necessarily knowledgeable. ICTs may sharpen some of the potential areas of tension already present in society as much as they may alleviate others or make them obsolete entirely.

Consumers

While many herald the developments set in motion by ICTs as consumers, the overall effect need not be that consumers benefit from ICTs and e-commerce in absolute as well as in relative terms (Dolfsma 2006). Certainly, goods offered are more likely to meet the preferences of consumers, surely those consumers in the 'long tail' of goods for which there is no big (local) market. Aggregation of previously unmet demand is much easier and less costly. In addition, new goods can be developed based on the preferences of consumers that can be gathered and processed relatively easily. Consumers will in effect increasingly co-produce goods

as they provide producers with extensive information about what it is that they would prefer to buy. Product differentiation is a strategy that firms are able to employ much more readily. Moving between suppliers is then less likely to appeal to consumers, however, as the new supplier will need to collect information afresh. Based on knowledge about the preferences of individuals, firms will also charge different prices to different (groups) of individuals, i.e. price discrimination as a tool is likely to become more prominently used. Will large groups of consumers with a variety of interests be able to organize themselves to fend off this possible development, sponsored by a relatively small group of devoted players with substantial resources?—Possibly, although not likely as per the argument of Mancur Olson in *The Logic of Collective Action* (1965).

Markets may thus become *less* transparent because of information overload: "can't see the forest for the trees". As information available grows the value of a reputation of a product or producer, paradoxically, perhaps, only increases (Faulhaber & Yao 1989).

Civilians

In some circumstances, some civilians are willing to express their views quite readily, possibly in the conviction that their voice will be anonymous and the repercussions of raising a voice will not come around. The play with identity that is possible in virtual environments gives many people the opportunity for personal development. The possible abuse of others will lead to calls to prevent anonymity.

Technical means to prevent (some) individuals from using the Internet anonymously to undermine existing relations of power and influence have been and will increasingly be implemented. Even without these, it is clear that Internet and the use of

the ICT infrastructure is not likely to undermine previously existing power relations in society (Jones 1998). Dominant individuals and groups can use their more extensive resources, including knowledge and skills, and networks to more effect. Within and between countries this has led to a discussion of the so-called 'digital divide'.

Civilians can legitimately raise their voice in democracies about their government and the policies it pursues. Electronic means of voting, including direct referenda, may increasingly become available. Will the use of such options lead to a more volatile political climate? This raises questions of a political philosophical nature – Shouldn't a government be installed based on a general policy platform, as conservative political philosopher Edmund Burke has argued, and held to account only after a number of years? It also raises questions of an economic nature. For instance: Can civilians be entrusted to provide a political climate and legal environment that is sufficiently stable and predictable for an economy to flourish? Nevertheless, if the internet and other ICT enabled technologies can be likened to the media / press, then a free press is likely to foster the economy as much as it does society and democracy (Sen 1999).

ICT & Globalization

The ICT sector is one of the major enablers for the globalization of society in general and the globalization of business in particular. Liberalization of market and international trade under the aegis of the WTO is another—indeed, these two factors might not be unrelated. Through the use of ICT, firms are able to interconnect their local branches into centralized IT systems. ICT also provides the means to cooperate effectively in geographically dispersed teams, for instance through videoconferencing and data sharing. One should not predict that any use of a

technology that is technically possible will also be used in practice. For instance, the conditions under which cooperation in virtual teams works best are vehemently researched into. Sometimes, work in virtual teams works well, most notably when the knowledge that team members need to have in common can be made explicit. Off-shoring then seems to offer great prospects. When the knowledge required is complex in some sense, and tacit, team members need to work closely together in close geographical proximity (Hansen 1999). While there is outsourcing of R&D, many high-tech industries are highly concentrated geographically to take advantage of such what might be called knowledge spill-overs.

At the same time, the production of much of the ICT hardware and software is shifting to Asia. The use of available ICT infrastructure and software allow for a range of services to be productively off-shored to countries where labour costs are low, countries in Asia and Eastern Europe, for instance. Call centres that offer services for which local or tacit knowledge about the customer plays a less important role are an example. Services that require communication between agents and timeliness, such as in the financial sectors for instance, are apt to make use of ICTs and globalize. Software development, a labour intensive activity, is increasingly outsourced too. Software being an information good, development of a piece of software can be easily monitored from a distance. From the discipline of economic geography it is to be expected that firms at first hardly embedded in the countries they have moved to in order to reduce costs may embed locally over time. Also, while the outsourcing of production facility, services and development is likely to be critiqued by a number of parties in the developed countries, it is likely to be a boon

for the developing countries at the receiving end.

Governance

Public Domain

Because of their strong economic and social benefits, ICTs receive a good deal of attention from governments. Even though the ICT sector itself involves only a relatively small part of a country's economy, the growth of this sector is generally believed to spur the development of the broader economy. Use of the products from this sector affects the entire economy; the financial sector has been the biggest investor in ICT hardware and software for many years in many countries.

Besides its clear benefits, ICT also has some drawbacks and poses new challenges to governments. First of all, ICT has a profound impact on government and the public sector itself (Heeks 1999). On-line government applications and services (e-government) enable information exchange between public institutions, but also between governments and citizens. Thus, e-government can contribute to the transparency of government and public decision-making. Furthermore, ICT can contribute to the democratic process in the form of on-line discussion groups for citizens. Other ICT opportunities involve personal identification, e.g., through the storage of biometrical data (finger prints, photographs, iris scans) on chips in passports.

Despite these advantages of information access and information sharing, governments are faced with the task to present on-line information appropriately and to prevent information overload and corruption. Other challenges to governments exist in terms of the 'digital divide' (OECD 2004b). While the most important issue has been to give also the elderly, the less-educated, and the poor access to ICT, recently the policy concern is shifting to the actual use of ICT, to prevent vulnerable

groups to further fall behind because they miss the benefits of ICT.

Other challenges for Western governments arise as a result of the aforementioned off-shoring of activities. In the short term this may result in job losses, but the long term challenge for governments is to retain and improve knowledge-intensive activities to prevent their knowledge economies from deteriorating. It is commonly held that in Western economies knowledge is the most important production factor to create additional value and therefore the basis for international competition.

To further stimulate the knowledge economy through the use of ICTs (see for instance the initiative of the European Commission "i2010: European Information Society 2010"), governments need to ensure the safety and security of (using) these technologies. Issues in this respect relate to the development of electronic signatures, the fight against spam, computer viruses, hacking. Other concerns involve the health and security implications of wireless technologies, such as the (alleged) radiation of UMTS (Universal Mobile Telecommunications System) base stations and regulation of hands-free driving. At the same time, however, such measures might restrict use of ICTs and the Internet, including use that might otherwise be beneficial.

One policy area that might have the effect of possibly overly restricting use of information is that of Intellectual Property Rights including patents and copyrights (Dolfma, 2005). Making copies without permission or payment of software programs, but of forms of content with a substantial entertainment value, such as music, movies and games as well is illegal. Enforcing existing IPRs more strictly or expanding scope and duration of IPRs might hamper exchange of knowledge – to the detriment of society at large (Mokyr 2002).

An overarching challenge for governments is that their policy is restricted to the confines of their territory. (Economic) activity that is outlawed in one country might be undertaken from another country. Policing the borders is much more difficult – though not impossible – if ICTs would not be present. The scope, therefore, for policy competition (e.g. tax havens, gambling, etc.) is broader. How likely is coordination of policy by (enough) governments across the globe?

Private domain

The strong network externalities for most ICT standards mean that ICT markets can ‘tip’ in favour of a *de facto* or *de jure* single standard (Shapiro & Varian 1999). If such a standard is based on proprietary technology, a monopolistic market structure results. Obviously, this raises concerns for the price and quality levels of the system. Typically, the installed customer base of a dominant standard faces considerable switching costs and coordination problems to adopt alternative technologies. Competing standards regularly coexist though and many monopolies are only temporary due to rapid technological progress. Some products obtain very dominant positions however, such as Microsoft’s operating system Windows. Such dominant positions raise policy concerns as ample opportunities exist for dominant firms in network industries to bundle products and oppose (potential) competition and deter innovation. Anti-trust or competition policy issues are likely to increasingly take on a cross border nature. A conflict with IP Law is also likely to be felt, as the standards tend to be protected under IPR.

These different market outcomes (e.g. coexisting standards and temporary monopolies versus sustained monopolies) confront policymakers responsible for competition policy with the difficult task to trade-off the static and dynamic efficiency of

markets. On the one hand policy makers want to ensure that consumers get a good deal in current markets and on the other hand they need to create the conditions to stimulate innovation of both incumbents and new entrants. While it is unclear if one should expect large, monopoly-like players or entrants and newly established firms to be more innovative, the tendency is to expect the latter to contribute to dynamic efficiency more, while standard economic theory indicates that monopoly-like players are likely to charge a high price and restrict output more than the public interest warrants.

One way governments influence the dominance of standards and competition in ICT markets is the creation and diffusion of open standards and technologies through standardization bodies. These institutions, such as the International Telecommunication Union (ITU) and the European Telecommunications Standards Institute (ETSI), bring together different nations and market participants for the creation of open standards. A successful European initiative for instance was the development of the GSM standard for digital mobile telecommunications. The combination of various hardware manufacturers and the adoption of this technology in all European countries created strong (supply-side) economies of scale and strong incentives for firms to develop complementary products (e.g., network elements and handsets).

A special case of government intervention involves the telecommunications market, which can be said to constitute a natural monopoly. In this market competition fails as the (last-mile) infrastructure is simply too costly to duplicate. However, technological developments have resulted in wireless alternatives, and starting in the 1980s public policy has been to regulate access to the telecommunications infrastructure to promote competition between service providers

(Laffont & Tirole, 2000). This process of deregulation telecommunications required heavy re-regulation (Emmons 2000) in terms of national regulators (such as the Federal Communications Commission, FCC, in the US) that aim to create a level-playing field, for instance by determining wholesale prices and taking care of dispute resolution between incumbents and new entrants.

Conclusion

As the technological paradigm underlying the fifth Kondratieff Long Wave in modern economic history, ICTs have a substantial and lasting impact on almost all aspects of economy and society. Despite the obvious advantages of technological progress and the adoption of ICT, this process confronts firms, individuals as well as governments with new challenges. New trade-offs need to be made, such as those related to the digital divide, ICT regulation, and ICT security.

Selected References

- Courvisanos, J. (2004) "Michał Kalecki as a Behavioural Economist: Implications for Modern Evolutionary Economic Analysis", in Z. Sadowski and A. Szeworski. (Editors) *Kalecki's Economics Today*. London: Routledge, pp. 27-41.
- Dolfsma, W. (2006b) "IPRs, Technological Development and Economic Development", *Journal of Economic Issues*, 40, 2.
- Dolfsma, W. (2006a) "Collective Consuming: Consumers as Subcontractors on Electronic Markets", *The Information Society*, 22, 3.
- Dolfsma, W. (2005) "Towards a Dynamic (Schumpeterian) Welfare Theory", *Research Policy*, 34, 1, 69-82.
- Dolfsma, W. (2004) *Institutional Economics and the Formation of Preferences*. Cheltenham: Edward Elgar.
- Dosi, G., C. Freeman; R. Nelson; G. Silverberg and L. Soete, (1988) (Editors), *Technical Change and Economic Theory*, London: Pinter.
- Emmons, W (2000) *The Evolving Bargain: Strategic Implications of Deregulation and Privatization*. Harvard Business School Press.
- Faulhaber, G.R. and D.A. Yao. (1989) "Fly-by-Night Firms and the Market for Product Reviews", *Journal of Industrial Economics*, 38, 65-77.
- Freeman, C. and C. Perez. (1988) "Structural Crises of Adjustment, Business Cycles and Investment Behaviour", in: G. Dosi, C. Freeman, R. Nelson, G. Silverberg and L. Soete. (Editors), *Technical Change and Economic Theory*. Pinter, pp. 38-66
- Hansen, M. (1999) "The Search-Transfer Problem: The Role of Weak Ties in Sharing Knowledge Across Organization Subunits", *Administrative Science Quarterly*, 44, 82-111.
- Heeks, R. (1999) *Reinventing Government in the Information Age*. London and New York: Routledge.
- Jones, S.G. (1998) (Editor) *Cybersociety 2.0—Revisiting Computer-Mediated Communication and Community*. Sage Publishing.
- Kleinknecht, A. and G. Van der Panne. (2008) "Long Waves", in: J. Davis and W. Dolfsma (Editors), *Handbook of Social Economics*. Cheltenham: Edward Elgar.
- Laffont, J. and J. Tirole. (2000) *Competition in Telecommunications*. Cambridge, MA: MIT Press.
- Mokyr, J. (2002) *Gifts of Athena – Historical Foundations of the Knowledge Economy*. New Jersey: Princeton University Press.
- OECD (2004a) *Recommendation of the Council on Broadband Development*. Paris: OECD.
- OECD (2004b) *Information Technology Outlook*. Paris; OECD.

- OECD (2005) *Communications Outlook*. Paris: OECD.
- Olson, M. (1965) *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, MA: Harvard University Press.
- Perez, C. (2002) *Technological Revolutions and Financial Capital: The Dynamics of Bubbles and Golden Ages*. Cheltenham, UK: Edward Elgar.
- Sen, A. (1999) *Development as Freedom*. Oxford: Oxford University Press.
- Shapiro, C. and H.R. Varian. (1999) *Information Rules: A Strategic Guide to the Network Economy*. Cambridge, Massachusetts: Harvard Business School Press.
- Shy, O. (2001) *The Economics of Network Industries*, Cambridge UP.
- Van den Ende, J. and W. Dolfsma (2005) "Technology Push, Demand Pull and the Shaping of Technological Paradigms: Patterns in the Development of Computing Technology", *Journal of Evolutionary Economics*, 15, 1.

Website

European Commission. *i2010—A European Information Society for Growth and Employment*. europa.eu.int/information_society/eeurope/i2010/index_en.htm

Wilfred Dolfsma

*Dept of Innovation Management & Strategy
University of Groningen, Groningen
The Netherlands
W.A.Dolfsma@rug.nl*

Ferdinand Jaspers

*Department of Entrepreneurship and
Innovation
Rotterdam School of Management (RSM)
Erasmus University Rotterdam
The Netherlands
E: fjaspers@rsm.nl*

Innovation Policy

Rachel Parker

Introduction

Innovation can be defined broadly to include the development and uptake of new technology, the introduction of new products, the utilisation of new market opportunities and the implementation of new business processes including new forms of work organisation or management structures and approaches. Innovation, or the commercial application of new knowledge, is of increasing importance to economic competitiveness given the growth in production and trade in high technology industries and knowledge intensive service sectors such as business services (Edquist, Hommen and McKelvey 2001).

An important field of innovation in modern economies is associated with the rapid development and application of information and communications technologies (ICTs). ICTs constitute an increasing share of value added, growth and employment and also impact on employment and productivity in other industry sectors. The structural transformation of modern economies associated with ICTs has led to an increase in the importance of information and knowledge resources (rather than physical capital) as inputs or factors of production.

Technology and product innovations are often given central attention in innovation research, however, organisational and managerial changes have been recognised as critical. Over the last two decades, understandings of the nature and process of innovation have advanced significantly. In the 1950s and 1960s, there was a view that innovation resulted from basic research, or in essence that scientific research acted as a 'push' for innovation. As such there was a great deal of emphasis on formal research and

development, undertaken either by governments or research and development units within business organisations. Radical innovations involving new products and new technological trajectories were thought to derive from basic research (Freeman 1995).

More recently, understandings of innovation widened to identify national systems of innovation (NSI). This approach is now one of the most influential contributions to current knowledge in the field. The NSI approach conceptualises innovation as a process of interactive learning between a firm and its environment, involving feedback mechanisms or loops. This involves complex interactions between a variety of institutions within the system as part of a continuous process involving incremental change, error and modification (Edquist 1997). Nosi, Saviotti, Bellon and Crow (1993:212), have defined the NSI as:

“the system of interacting private and public firms (either large or small), universities, and government agencies aiming at the production of science and technology within national borders. Interaction among these units may be technical, commercial, legal, social and financial, in as much as the goal of the interaction is the development, protection, financing or regulation of new science and technology.”

State and Innovation Policy

There are several perspectives on the role of the state in promoting economic innovation. One approach emphasises market rewards and incentives, rather than social relations and non-market institutions, as the basis of policies designed to promote innovative behaviour. The market based approach views innovation as a market process driven by entrepreneurs who are motivated by the possibility of earning high rewards for

undertaking potentially high growth activities. In this approach, the role of the government is to ensure that there are appropriate framework conditions for entrepreneurial activity including adequate potential for earning high rewards for high risk activity and tax and income incentives for self employment, risk and initiative. This approach is consistent with a policy emphasis on small business and entrepreneurship and typically involves policies that are designed to reinstate market incentives for entrepreneurial and innovative activity that were said to be eroded by large governments, high taxation or general welfare provision (Verheul et al 2002).

A further dimension of the innovation literature with important implications for public policy is that which contrasts the role of the state in promoting innovation in the knowledge economy with its role in the initial post World War II era, in which the state targeted 'national champions' or key large corporations in important industry sectors. Audretsch and Thurik (2001) identify four characteristics of the role of government in the knowledge economy that can be contrasted with its role in the managed economy. First, governments have shifted their emphasis from regulation (through public ownership and anti-monopoly) to stimulation by creating an environment of knowledge creation through research policies and education policies. Second, governments focus on inputs into the innovation system (such as skills and knowledge) rather than outputs such as favoured firms or industries. Third, there is an increased emphasis on regional and local policies in the knowledge economy rather than national policies given the importance of local learning processes and tacit knowledge. Fourth, there is a public policy emphasis on high-risk venture capital finance (rather than low risk capital), which is

essential to the commercialisation of new ideas.

As explained above, the NSI approach has drawn attention to the process of innovation and the range of institutions that make up the innovation system. Research on innovation systems has focused on both national and regional levels of analysis and has improved our understanding of the importance of local forms of economic governance in a global era. The NSI approach recognises the importance of institutions and spatial proximity in the process of learning and knowledge creation and as such has emphasised the continuing relevance of national and regional governance systems in influencing innovation, industry development, economic change and economic growth (Morgan 1997). The innovative activities of firms are affected by their spatial proximity to knowledge sources that are external to the firm including universities, public research institutions and other firms, especially knowledge intensive firms (Lundvall 1992).

The importance of spatial proximity arises from learning by interacting which results in the transfer of local knowledge including tacit knowledge resulting from learning by doing and using (Asheim and Clark 2001). Given the spatial dimension of knowledge transfer, urban and regional policies are also of particular relevance to innovation policy. Regional innovation policy involves the promotion of learning processes amongst a variety of actors including firms, local government authorities, public research institutions and education and training bodies (Asheim and Isaksen 1997).

The NSI approach has also explored the role of public policy in influencing the science and technology infrastructure as well as the skills level of the labour force as critical influences over innovation. As Lundvall (1992: 14) has highlighted, "the public sector

plays an important role in the process of innovation ... it is involved in direct support of science and development, its regulations and standards influence the rate and direction of innovation, and it is the single most important user of innovations developed in the private sector.”

The state can influence innovation by making direct contributions to technology development through public funding of research institutions and the science base. It has been a demanding user of new technologies, creating initial markets and contributing to product development and innovation. Public policy initiatives thus encourage interactive processes between actors within the innovation system. Given the importance of intangible capital, particularly in the form of tacit knowledge, the state's promotion of the exchange of information and ideas through the coordination and transfer of skills and knowledge between organisations is critical (Edquist, Hommen and McKelvey 2001).

Comparative Perspective

Examples of policies to promote innovation can be taken from different national contexts. In the Danish ‘negotiated economy’, the state has engaged in dialogue, communication and negotiation with industry. Policies affecting business innovation in Denmark have been developed and coordinated in local communities by a multitude of agents including industrial firms, banks, and public institutions at various levels of government throughout the 1990s (Drejer, Kristensen, Laursen 1999). As such, the state has played a key role in coordinating dialogue and interaction amongst key economic actors enhancing the process of information sharing and interactive learning.

At the European level, the European Union has focused its attention on the need to redress technological divergence between

regions. There is concern about variation in level of research and development expenditure and innovative activity between regions in the European Union, which constitutes one of the most significant forms of economic divergence within the Union (Guersent 2001). Structural funds have been used to invest in new technologies and access to information-society resources including telecommunications networks. A focus on the information-society constitutes one aspect of the shift in the European Union away from sectoral policies, focused principally on agriculture, towards policies oriented to regional development, particularly technological development within regions (Buunk, Hetsen and Jansen 1999).

In the United States, the Defence Department has played a role in technology development through its support of defence technology and weapons manufacturing, with significant implications for industries such as aeronautics and microelectronics. The state has played a critical role in funding technology development as well as acting as the major user (in some cases sole user) of key new technologies including the internet, transistors, computers and information systems in their early stages of development (Chomsky 1998). As such, the state has played a role in funding and using key technologies. In Japan, the state has massively increased its funding of science and technology in the 1990s, despite mounting fiscal pressure. The state has prioritised science and technology policy by coordinating the government's overall policy framework through the cabinet office. Further, the mobility of researchers between universities and industry has been emphasised, as has the commercial relevance of academic research (Sato 2001). Public policy initiatives have therefore focused both on developing the knowledge base in science and technology and enhancing the

interactions between universities and industry for the purpose of knowledge transfer.

In Singapore, the state has sought to transform the industrial economy towards high value-added activities by targeting high value-added manufacturing, particularly the electronics industry and services such as business services and R&D. This involved the promotion of producing firms and supporting and components industries. The strategies to achieve this goal include the establishment of industrial parks, providing venture capital and business management support to key local enterprises and co-investment activities designed to reduce business risk (van Grunsven & van Egeraat 1999:151-152). The state has therefore sought to develop clusters of economic activity around key related industries and services.

Innovation System and Industrial Competitiveness

The systems approach to innovation has explained that a nation's values and norms, state institutions, industrial relations system, finance sector, industry associations, trade unions and business organisations constitute a system within which innovation takes place. It is common to draw a broad distinction between two distinctive types of business systems or models of capitalism—the coordinated and competitive types—which have different implications for innovation. Each type of capitalism is associated with different processes of learning, knowledge transfer and innovation. Particular systems are thought to give rise to competitiveness in different types of industries. In respect of innovation, coordinated and competitive models are thought to give rise to quite different innovative capacities (Casper 2000, Haake 2002, Soskice 1999, Whitley 2000).

The model of co-ordinated capitalism is characterised by well-developed forms of corporate organisation, often involving

specialisation on core competencies, close linkages between industry and the finance sector resulting from cross-ownership and control between enterprises, long-term stable relationships with customers and suppliers and particular forms of inter-firm co-operation in relation to information sharing and the pooling of resources for research and development, design and marketing. In addition, the industrial relations system in the co-ordinated model is characterised by collective bargaining and labour market programs and institutions that emphasise skills development in the workforce and security of tenure. Business associations tend to be encompassing and well integrated with state policy making institutions. Culture is oriented towards cooperation, trust and equality.

These arrangements facilitate the sharing of information between economic actors and allow for the pooling of resources for research, marketing and information gathering. They also tie economic actors to existing relationships, either outside the firm (with other institutions such as research and education institutions) or within the firm (between managers and employees). As such, incremental innovations, or gradual improvements in existing products and production processes are favoured, rather than more radical innovations involving new technological trajectories which would require firms to change the composition of their workforce or to disrupt long term relationships outside the firm.

The coordinated model, typical of Germany and Sweden, is often thought to favour incremental innovations, particularly in medium-technology industries. The German and Swedish models have proved to be very successful in the diffusion of incremental innovations within existing industrial enterprises and have been less noted for radical innovations in new industries. The

focus of innovation has been on the development and application of new technologies to existing production activities, as opposed to the development of new products and processes. Further, the relatively high business costs and rigid labour and social regulation in Sweden and Germany have been regarded as driving firms to adopt a strategy of diversified quality production that has involved the mass production of high quality products rather than focusing on cost minimisation as the basis of competitiveness (Streeck 1997).

However, it should be noted that both Germany and Sweden are performing well in key new economy activities, including information and communications technologies and biotechnology. The functional flexibility of the Swedish and German labour forces as well as sectoral and regional innovation policies may explain their capacity to move into new industry sectors (Asheim and Clark 2001, Casper 2000). In particular, it has been explained that the Swedish coordinated system, which involves high levels of welfare provision and labour organisation, has been compatible with entrepreneurial activity in knowledge intensive sectors because of the background of social protection. Potential entrepreneurs are willing to undertake high-risk activities given the knowledge that society will provide fundamental support in the case of failure. Unions have also been supportive of economic transformation and the adoption of new technologies despite their potential impact on unemployment given the protection of the welfare system (Benner 2003).

The competitive model is instead characterised by weakly organised business groups and unions, decentralised determination of wages (at the level of enterprises), a highly competitive labour market with high-labour turnover, a financial system heavily dependent on capital markets

providing ready access to high-risk capital, hierarchical M-form diversified firms not typically categorised by decentralised inter-organisational relationships or participation in clusters, a strong emphasis on competition and anti-trust, and an unwillingness of the state to interfere with the investment and production decisions of private firms. In this model, as firms are less embedded in relationships with other firms and institutions, they are less tied to existing production strategies or product lines. As such, they are thought to be more likely to undertake radical innovations, which result in the destruction of existing competencies and the acquisition of new organisational capabilities based on new skills and technological competencies. The competitive system is also regarded as encouraging individual initiative and risk taking and therefore favouring new firm start-ups in new and rapidly changing industry sectors.

The competitive model, typical of the United States, is regarded as conducive to the rapid development of new technologies. In the United States, large venture capital markets, an entrepreneurial business culture, close linkages between universities and industry and a highly mobile labour force have been regarded as critical elements of the business system which have encouraged the development of new technology firms in new industries. As Mowery and Rosenberg (1992:48) explain:

“the successive waves of new product technologies that have swept through the postwar U.S. economy, including semiconductors, computers, and biotechnology, have been commercialized in large part through the efforts of new firms. The role of small firms in commercializing new technologies in the United States during this period appears to contrast with the pattern in both Japan and Western Europe, where established firms

in electronics, pharmaceuticals, and other industries have played a more significant role in technology development.”

The highly flexible external labour market in the U.S. may have constrained investment in training within firms, but it is thought to have facilitated the success of regional clusters of high technology firms, such as in Silicon Valley, by enabling technology diffusion between firms by highly skilled employees and also by providing the possibility for researchers in public industrial laboratories and universities to move in to industry and commercialise their innovations (Saxenian 1996). In addition, the well developed venture capital market in the U.S., which has been willing to support risky new ventures, is regarded as having contributed to the development of new small firms in new industries, such as biotechnology.

It is therefore possible to draw a distinction between competitive and coordinated market economies and to highlight their different implications for innovation. The former have an emphasis on market relations and autonomous business units. In the latter, relations within corporations (between management and employees) and with external organisations and institutions (including other firms, suppliers, customers, research institutions and educational institutions) are more developed.

It is difficult to draw any definite conclusion about the relative innovative performance of competitive or coordinated economies. It would seem that some coordinated economies, such as Sweden, have been successful in a range of industries including knowledge intensive activities such as medium technology manufacturing industries and more recently in ICT. However, it is also the case that the competitive economy of the USA has been successful in knowledge intensive activities, particularly in rapidly

changing sectors and in the finance and insurance service sector.

Social Implications

It should be noted that not all knowledge intensive activities are alike and while innovation policy can encourage participation in the knowledge economy, which has the potential to contribute to rising wages and profits, it does not necessarily result in wide spread improvement in work and employment arrangements, particularly in competitive systems in which labour markets are relatively flexible and there is a high level of inequality in wages and hierarchical forms of business organisation (Asheim & Clark 2001:809). As such, innovation policy has significant implications for the future of work and needs to be developed in conjunction with labour market policies.

The challenge for innovation policy is to identify strategies that ensure that high-skilled and knowledge intensive activities have broad social benefits and are not isolated in pockets of highly profitable and growth oriented firms and regions. As such, innovation policy should be considered in conjunction with issues associated with *the distribution of income and wealth*. Research on types of capitalism and innovation has shown that innovation occurs in different ways in the competitive and coordinated institutional environment. However, in the case of coordinated economies, it would seem that success in knowledge industries is more broadly felt throughout society as a consequence of a high level of income equality and social cohesion. So while the competitive and coordinated models both have strengths and weaknesses in their capacity to encourage technology development and innovation in different kinds of industries, it is in the coordinated model that the rewards of the new economy

translate into widely experienced social benefits (Lundvall 2002).

Selected References

- Asheim, B. and A. Isaksen. (1997) "Location, Agglomeration and Innovation: Towards Regional Innovation in Norway?", *European Planning Studies*, 5, 3, 299-330.
- Asheim, B. and E. Clark (2001) "Creativity and Cost in Urban and Regional Development in the 'New Economy'", *European Planning Studies*, 9, 7, 805-811.
- Audretsch, D. and R. Thurik (2001) "What's New about the New Economy? Sources of Growth in the Managed and Entrepreneurial Economies", *Industrial and Corporate Change*, 10: 267-315.
- Benner, M. (2003) "The Scandinavian Challenge: The Future of Advanced Welfare States in the Knowledge Economy", *Acta Sociologica*, 46, 2.
- Buuk, W., H. Hetsen, A. Jansen (1999) "From Sectoral to Regional Policies: A First Step Towards Spatial Planning in the European Union?" *European Planning Studies*, 7, 1, 81-99.
- Casper, S. (2000) "Institutional Adaptiveness, Technology Policy, and the Diffusion of New Business Models: The Case of German Biotechnology", *Organization Studies*, 21, 5, 887-914.
- Chomsky, N. (1998) "Power in the Global Era", *New Left Review* 230: 3-27.
- Drejer, I. and F. S. Kristensen and K. Laursen (1999). "Cluster Studies as the Basis for Industrial Policy: The Case of Denmark", *Industry and Innovation*, 6, 2, 171-190.
- Edquist, C. (1997) "Systems of Innovation Approaches—Their Emergence and Characteristics", in C. Edquist (Editor), *Systems of Innovation: Technologies, Institutions and Organisations*. London: Pinter: 1-35.
- Edquist, C., L. Hommen and M. McKelvey. (2001) *Innovation and Employment: Process versus Product Innovation*. Cheltenham: Edward Elgar.
- Freeman, C. (1995) "The 'National System of Innovation' in Historical Perspective", *Cambridge Journal of Economics*, 19, 5-24.
- Guersent, O. (2001) "The Regional Policy of the European Union: A Balance and Outlook", *Regional Studies*, 35, 2, 163-175.
- Haake, S. (2002) "National Business Systems and Industry-Specific Competitiveness", *Organization Studies*, 23, 5, 711-736.
- Lundvall, B. (1992) (Editor) *National Systems of Innovation: Towards a Theory of Interactive Learning*. London, Pinter.
- Lundvall, B. (2002) *Innovation, Growth, and Social Cohesion: The Danish Model*. Northampton, MA: Edward Elgar.
- Morgan, K. (1997) "The Learning Region: Institutions, Innovation and Regional Renewal", *Regional Studies*, 31, 491-503.
- Mowery, D. C. and N. Rosenberg (1992) "The U.S. National Innovation System", in R. Nelson (Editor), *National Innovation Systems: A Comparative Analysis*. New York: Oxford University Press.
- Niosi, J., P. Saviotti, B. Bellon and M. Crow (1993) "National Systems of Innovation: In Search of a Workable Concept", *Technology in Society*, 15, 2, 207-227.
- Sato, Y. (2001) "The Structure and Perspective of Science and Technology Policy in Japan", in P. Larédo and P. Mustar (editors), *Research and Innovation Policies in the New Global Economy*. Cheltenham: Edward Elgar, 79-114.
- Saxenian, A. (1996) Beyond Boundaries: Open Labour Markets and Learning in Silicon Valley. *The Boundaryless Career: A New Employment Principle for a New Organisational Era*. In M. B. Arthur and D. M. Rousseau Eds., New York, New York University Press.
- Soskice, D. (1999) "Divergent Production

- Regimes: Coordinated and Uncoordinated Market Economies in the 1980s and 1990s”, in H. Kitschelt, P. Lange, G. Marks and J. D. Stephens (Editors), *Continuity and Change in Contemporary Capitalism*. London: Cambridge University Press.
- Streeck, W. (1997) “German Capitalism: Does it Exist? Can it Survive?” *New Political Economy*, 2, 2, 237-256.
- Van Grunsven, L. and C. Van Egeraat (1999) “Achievements of the Industrial 'High Road' and Clustering Strategies in Singapore and Their Relevance to European Peripheral Economies”, *European Planning Studies*, 7, 2, 145-173.
- Verheul, Ingrid; Sander Wennekers; David Audretsch and Roy Thurik (2002) “An Eclectic Theory of Entrepreneurship: Policies, Institutions and Culture”, in D. Audretsch; Roy Thurik; Ingrid Verheul and Sander Wennekers (Editors), *Entrepreneurship: Determinants and Policy in a European-US Comparison*. London: Kluwer Academic Publishers.
- Whitley, R. (2000) *Divergent Capitalisms: The Social Structuring and Change of Business Systems*. Oxford: Oxford University Press.

Websites

- DRUID. (Danish Research Unit for Industrial Dynamics.) www.druid.dk
- CORDIS Guidance and Support. www.cordis.lu/innovation-smes
- OECD. www.oecd.org

Rachel Parker
Brisbane Graduate School of Business
Queensland University of Technology
Brisbane, Queensland
Australia
r.parker@qut.edu.au

Interlocking Directorships

Bruce Cronin

Introduction

When a director of one company at the same time serves on the board of another company, the two companies are said to be interlocked by that director. Through this linkage each company has potential access to information about the activities of the other, either explicitly as intelligence transferred by the director or implicitly in shaping the director's perspective and general views. Director interlocks formed by executive directors, employed by the firm, are generally interpreted as more instrumental for the firm than those formed by non-executive directors. Firms are usually interlocked with more than one other company and often form a web of relationships involving many businesses.

History

The first studies of interlocking directorships were undertaken by the US government in the early 1900s during investigations into collusion in the railroad and banking industries. The Pujo Commission mapped the director interlocks among the principal US banks and finance companies as part of their identification of a 'Money Trust' around J. P. Morgan & Co. (Pujo Commission, 1913). The interlocks were widely interpreted as a mechanism by which the companies operated a cartel, setting prices and regulating markets among the members of the trust (e.g. Brandeis 1967) and on a populist wave the 1914 Clayton Act outlawed director interlocks among competing firms.

Directors themselves tend to be dismissive of the potential for interlocks to provide coordinating or collusive benefits for firms, seeing directors generally having little influence over the operational level at which price setting and other market activities take

place. The phenomena of interlinks is not seen as evidence of collusion but merely reflecting the limited supply of skilled and experienced candidates. However, utilising methodological advances in the field of social network analysis, the study of interlocking directorships has become increasingly sophisticated, allowing the identification of persistent patterns of interaction that go beyond supply shortages or mere chance.

Further, instances of large scale collusive and fraudulent behaviour among leading US and UK corporations in the late 1980s and also in the late 1990s and early 2000s have led to a renewed review of the independence of external directors by capital market regulators, drawing on some of this research. In 2003 the New York Stock Exchange amended its listing rules to require that a majority of board members have no 'material relationship' beyond share ownership in the company, including having served as an employee of a commercial partner or advisor during the previous three years (NYSE 2003:4).

Models of Corporate Governance

Concern about the collusive potential of director interlocks resides primarily in the US and UK, where the board of directors is seen as an important counterbalance to the personal interests of managers. In these countries, directors, elected by a firm's shareholders to represent their interests, establish broad parameters for the day to day activities of the firm's management, in an effort to minimise opportunist behaviour by the latter for personal gain. The ability of directors to scrutinise managerial activity has thus been increasingly seen as critical in the light of poor performance and instances of fraud and the degree of participation of executive directors in board activities has come into question. While there has been a growing emphasis on the importance of non-

executive or 'outside' directors in ensuring oversight independent of management, such scrutiny is seen to be compromised when directors hold a material interest in the firm, such as that provided by interlocks with interested parties.

In the UK there is little statutory prescription regarding the governance structure of firms, beyond the requirement for any firm to have two directors and to report to shareholders annually on their stewardship. But the general form of governance of firms listed on the UK Stock Exchange is the representation of shareholders by a board of 8-10 directors, chaired by a non-executive director but including the chief executive and 3-4 other executive directors. Senior managers additionally normally meet separately and more frequently as an executive committee. Since the Cadbury Report (1992) on corporate governance there has been increased emphasis on the role of non-executive directors on boards and more explicitly defined governance tasks (Conyon 1994).

United States legislation is more prescriptive, with the Securities and Exchange Commission regulating the listing rules of stock exchanges and enforcing corporate disclosure about governance arrangements. The boards of the largest US firms tend to have 9-15 members. As with the UK, boards combine inside and external directors but in the US executive directors are a minority, typically 30%, reflecting the more prescriptive regulation of governance arrangements. Only around half of these external directors could be classified as independent, however. Further, unlike the UK, boards are normally chaired by the chief executive or a former executive (Coles & Hesterly 2000).

The counterpoising of shareholders and managers is less stark in the governance structures of Continental Europe where a

range of stakeholders is normally represented on corporate supervisory boards and executives are statutorily excluded. In particular, reflecting the importance of institutional capital funding, banks and large institutional shareholders are typically directly represented on German and French boards. French chief executives are widely represented on many other firms, something as a duty to the general managerial corps. Large firms in both countries have statutory representation from the workforce on the supervisory board. Thus, by integrating a range of stakeholder perspectives into the governance structure in Continental Europe the issue of director independence from management tends to be of less concern; the emphasis is on an institutional 'balance' between managerial and supervisory boards (McCarthy & Puffer 2002).

A similar concern to balance stakeholder interests characterises the Japanese model of corporate governance. Not just banks, but a web of businesses with mutual interrelationships, are represented in the governing consultations of *keiretsu* business groups. Further, government agencies are also represented through the *amakudari* system where retired officials are employed in private sector managerial positions. Consequently Japanese boards have been typically large, with an average size of 30, and independent outside directors are rare. While *keiretsu* ties weakened with the crisis of the Japanese banking system and growing foreign acquisitions through the 1990s, changes to the governance system towards the US model, such as the 2002 Commercial Code, were largely cosmetic, power remaining with the personal network around the corporate president (Ahmadjian 2000).

Structures of Capitalisms

These national differences in governance structure are reflected in the considerable

body of research from the 1970s that has identified distinct structures of director interlocking in different countries. Together, these differences support the notion of distinctive 'national capitalisms'.

In a systematic cross-national comparison of the concentration of director interlocks Stockman, Ziegler and Scott (1985) found the number of interlocks per firm lowest in the UK (4.7), higher in the US (10.46), and highest in continental Europe (12.36). Windloff (2002) found a more pronounced pattern in multiple interlocks with another firm, lowest in the US (0.6%) and UK (2.1%) and much higher in continental Europe (14-23%). Recent investigation into cross-national directorate interlocks has found only minor variations on this pattern, with the identification of discrete 'Atlantic', European and Japanese networks (Carroll & Fennema 2002).

Scott (1991) argues that the distinct Anglo-American interlock structure reflects the greater reliance on financial institutions for capital funding in these countries. By contrast, the greater concentration of interlocks in Europe reflects the closer institutional arrangements between banks or investment companies and industry there. Similarly the directorate structure of Japanese *keiretsu* or South Korean *chaebol* can be related to their distinctive capital funding structures. However, restricting director relationships to a matter of capital funding is somewhat reductionist. In fact, studies of the interlock structure in the Anglo-American semi-periphery (Canada and Australasia) have found much greater interlock concentration there despite similar capital market structures to the US and UK (Ornstein 1989; Alexander 1994).

Moreover, at a firm level, in the US at least, little evidence of a relationship between capital dependence and interlocks with financial firms has been found (Mizruchi &

Stearns 1988). Rather, director interlocks vary with other firm characteristics, interlocking greatest in larger firms, financial institutions, firms with major minority shareholders, and domestic rather than foreign firms (Dooley 1969; Ornstein 1984; Carroll & Armstrong 1999).

In general, studies of director networks have identified a unitary structure within a country, with secondary cliques around regions (in the US) or financial institutions (Dooley 1969; Sonquist & Koenig 1975; Mariolis 1975; Mintz & Schwartz 1981; Mizruchi 1982).

What do Interlocks Do?

The populist legacy of antipathy to cartels informed initial attempts to analyse the managerial implications of director interlocking, interpreting these relationships as mechanisms of collusion to various degrees. Mace (1971) pointed to the potential regulatory transgressions arising from the flow of inside information from board to board, while directors themselves insisted they erected Chinese walls against any conflict of interest. Early research focused on the influence of major family-owned corporate groupings, such as Morgan and Rockefeller (Domhoff 1967, Zeitlin 1974), including ties and interchanges with government personnel (Freitag 1975). These institutional overlaps were seen to provide the basis of the shared ideological outlook and cohesive action of a capitalist social class. Yet for every case of business cohesion, there has been little shortage of cases of business division and disunity (see Mizruchi's 1992 survey).

Less instrumental variations of interlock research have concentrated on resource advantages accompanying interlocking. Indirect interlocking with competitors has been found to increase in times of industry uncertainty, for example (Lang & Lockhart

1990). Financial institutions, in particular, have been seen to constitute significant intersections within the interlock networks. These would recruit directors from major companies to their boards to assist their general business intelligence (Baum & Stiles 1965; Mintz & Schwartz 1981). Some studies suggest a tighter relationship, finding that banks tend to draw their directors from the companies they lend to (Bearden 1987) or arguing that banks place representatives on the boards of companies they lend to as a means of closely supervising their investments (Sweezy 1953, Kotz 1978). Some, following Hilferding's suggestion, argue that concentrated director interlocks between banks and industrial capitalists represent a distinct form of business, finance capital (Fennema & Schijf 1979; Overbeek 1990; Carroll, 1986). On balance, however, such finance-centred networks appear to be fluid, representing a 'polyarchic' rather than 'oligarchic' financial hegemony (Mintz & Schwartz 1985, Scott, 1985). Direct resource-exchange activities seem able to account for at best a small minority of directorate interlocks.

Less conspiratorial accounts are now more prominent, with interlocks seen to provide directors with a 'scan' of the business environment and business practices. The process is evident in the following accounts by non-executive directors reported by McNulty and Pettigrew (1999:54,63):

- '[Name of country] I knew well ... it has enormous potential. I said "go there, acquire a good team of people and you will get in at a price which is sensible and attractive". That they have done.'
- '[On joining the firm] one of the first things I said was "what about strategy and plans". At [name of another company] we have ten-year, three-year and one-year plans ... they do not have that ... So next week at [name of company] we are going away for

two days to a hotel down in the country and we are having senior executives put out as close to their first shot of a plan.'

The scan, the breadth of current experience, is a major reason in the appointment of external directors (Useem 1984). Information gained through this scan is given priority by directors because the sources are familiar, a major factor in social learning (Bandura 1986; Galaskiewicz & Wasserman 1989). The direct contact also allows intimate knowledge (Davis 1991; Meyer 1994). Directors who have been parties to the adoption of a practice in one firm may become committed advocates of this in other firms (Palmer *et al* 1993). However, direct contact is not always necessary; structurally equivalent networks are likely to expose participants to similar problems and solutions (Burt 1980).

On one level, the business scan is seen by some as providing the basis for an important governance function for the economy as a whole. Directors at the centre of this broad information network are seen to form a core 'inner circle' with privileged influence on decisions on economy-wide capital allocation and regulation (Mintz & Schwartz 1985; Useem 1984). While this brings us back to the realm of elite collusion, to some extent, detailed investigations into the 'inner core' have found a much more dynamic process of alliance, defection and social learning, underpinning increased board activism, for example (Westpahl & Zajac 1997). Thus, examination of the role of the business scan made possible by director interlocks has concentrated increasingly on the individual firm level.

Yet evidence of a systematic relationship between director interlocks and profitability has been elusive (Fligstein & Brantley 1992), suggesting that the potential information channels identified are used in only a limited manner. In part, this finding may reflect

methodological limitations; research in this area is almost universally cross-sectional, while social learning is a longitudinal process. For example, director interlocking has been found to be a frequent *response* of firms to financial difficulty (Richardson 1987; Mizruchi & Stearns 1998), so a simple cross-sectional association between interlocks and profitability is unlikely to be found. Similarly, the *content* of information passed through interlocks may change over time (Westphal *et al* 1997). And the *context* of the information may change, as when alternative information sources are available such as business media coverage of an organisational practice (Haunschild & Beckman 1998). Few studies examine the specific mechanics of the transfer of information through interlocks (Mizruchi 1996).

Comprehensive investigation of network effects is also hindered by the wide variety of interlocking that takes place among directors. Interlocks arise in many ways and do not necessarily simply constitute an instrument for the firms involved (Ornstein, 1984). Some interlocks may be intentional, aimed to secure specific relationships with other firms, resource-associated 'strong ties'. In other cases, the intentionality may be less clear, as when a 'professional' director is recruited because of their broad links with the general business community, scanning-associated 'weak ties'. Other links may arise accidentally, when a director is selected for their experience or acumen.

Longitudinal studies of director interlocks between pairs of firms broken by retirement or death have found only around 15% subsequently replaced, undermining suggestions of collusive or resource dependent behaviour in this activity (Koenig, Gogel & Sonquist 1979; Palmer, 1983; Stearns & Mizruchi 1986), although there is some association with profitability among such ties (Richardson 1987). However,

around half of broken ties are reconstituted with similar *types* of firms, which, while not associated with profitability, does support the notion of interlocks providing channels of broad business intelligence.

Despite the conceptual challenges of identifying the mechanisms of these inter-firm relationships and their embryonic status, there is growing evidence of the effect of these intelligence channels on the strategies and governance of major firms. A variety of studies have now found the adoption of a range of business practices associated with director interlocking.

Geletkanycz and Hambrick (1997) identified a relationship between director interlocking and a firm's conformity with the typical strategy in an industry in terms of resource allocation such as capital, advertising, research and development and overhead spending and gearing. The greater the number of directors from outside a firm's industry, the more divergent the firm's strategy from the rest of the industry in these terms. Westphal *et al.* (2001) also found an indirect effect in these terms. The more interlocked firms conformed with the resource allocation norms of an industry, the more the focal firm did as well.

Haunschild (1993) found managers imitated the corporate acquisition behaviour of firms they served on the boards, particularly if these were firms in similar industries or banks (Haunschild & Beckman 1994). Further, premiums paid for acquisitions were similar among firms sharing directors as well as among those using the same investment banker (Haunschild 1994). Again, Westphal *et al* (2001) found a secondary effect of director interlocks on acquisitions. The more interlocked firms imitated the normal acquisition pattern in an industry, the more the focal firm did as well.

Separate studies have identified the spread of a number of tactical defences to the wave of hostile takeover attempts in the 1980s via the interlocking directorate in the United States. Davis (1991) examined diffusion of the poison-pill defence, where managers issue an option for shareholders to purchase shares at a great discount in the event of a takeover without board approval, thus greatly increasing the cost of the takeover. Firms sharing directors with firms that had adopted the tactic were more likely to adopt it themselves. Wade *et al.* (1990) found the incidence of golden parachutes, where managers receive large compensation payments in the event of a takeover and thus increasing the vigour of defence, positively associated with the number of boards a CEO served on. In Davis' (1991) study, however, where interlocking directors had a material interest in the firm, and thus were damaged by the reduction in shareholder value, a constraining effect was evident. Similarly, greenmail, a firm repurchasing its own stock at an above-market price, was found less likely where director interlocks involved a material benefit for the director (Kosnik 1987). So, different tactics appear to be diffused through different interests in director networks.

The diffusion of business practices through director networks extends to the very organisational structure of firms themselves. Alongside economic influences, firms sharing directors with firms using a multidivisional structure have been found more likely to adopt the same form themselves (Palmer *et al.* 1993; Fligstein 1985). Similarly, Mizruchi & Stearns (1994) found large US firms borrowing a greater proportion of funds when sharing directors with financial institutions. They speculated this may reflect greater access of these firms to information or advice on funding or greater confidence by lenders in firms they have greater knowledge of.

More specific business practices also appear to diffuse through director networks. Chua and Petty (1999) found Australian firms more likely to adopt ISO quality accreditation if they shared directors with firms that already had this accreditation. While Westphal and Zajack (1997) found that firms did not directly imitate the compensation policies of interlocked firms, O'Reilly *et al.* (1988) found CEO salaries strongly associated with the average salaries of the external directors on the compensation committee and Westphal *et al.* (2001) found an indirect effect. The more interlocked firms conformed with the compensation norms of an industry, so too did the focal firm.

The activities of business in society are also influenced by director interlocks, at least in the US. Galaskiewicz & Wasserman (1989) found the pattern of corporate charitable donations associated with the donations of interlocked firms. Mizruchi (1992) found corporate contributions to political campaigns more strongly associated with director interlocks, particularly with financial institutions, than particular interests of firms. Burris (1991) found executives with multiple directorships more likely to donate to Republican candidates but firms with more director interlocks more likely to contribute to Democrats, firm contributions more influenced by defence contracts and issues of regulation (Burris 1991). More interlocked firms, particularly those at the centre of the interlocking directorate network (Mintz 1995) also commit more resources to lobbying government.

Thus, the network of interlocking directors appears to play an important role in spreading business practices from firm to firm and generalising specific practices as industry norms. While information on these practices is readily available in the business media and professional forums, the trusted or insider character of directors appears to add some

legitimacy or perhaps privileged knowledge of their applicability.

Given the limited explicit acknowledgement of this process by board members, as shown by Useem's (1984) interviews for example, this method of diffusing business practices is unlikely to be professionally rigorous. Business practices are more likely to spread through this channel by chance, whim or bandwagon than by measured evaluation of alternatives. This suggests that the channel may represent a significant weakness in the governance function of boards of directors.

Thus, while in governance terms great store is relied on external or non-executive directors as a countervailing influence to internal managerial interests on boards, the study of interlocking directorships suggests the independence of these figures is overstated. Not only are external directors recruited from a limited social and managerial circles as the executive directors, the interlocking underpins an often homogenising community, whether in the extreme of links to an 'inner circle', or more broadly as a conduit for trends in business practice.

Selected References

- Ahmadjian, Christina L. (2000) "Changing Japanese Corporate Governance", *Japanese Economy*, 28, 6, 59-84.
- Alexander, M. (1994) "Business Power in Australia: The Concentration of Company Directorship Holding Among the Top 250 Corporations", *Australian Journal of Political Science*, 29, 40-61.
- Bandura, A. (1986) *Social Foundations of Thought and Action*. Englewood Cliffs, NJ.: Prentice-Hall.
- Baum, D. and N. Stiles. (1965) *The Silent Partners: Institutional Investors and Corporate Control*. Syracuse, NY: Syracuse University Press.
- Beardon, J. (1987) "Financial Hegemony, Social Capital and Bank Boards of Directors", in M. Schwartz (Editor), *The Structure of Power in America*. New York: Holmes and Meier, 48-59.
- Brandeis, Louis D. (1967) *Other Peoples' Money and How Bankers Use It*. Edited by Abrams. New York: R. M. Harper & Row.
- Burris, V. (1991) "Director Interlocks and the Political Behavior of Corporations and Corporate Elites", *Social Science Quarterly*, 72, 537-551.
- Burt, R.S. (1980) "Cooptive Corporate Actor Networks: A Reconsideration of Interlocking Directorates Involving American Manufacturing", *Administrative Science Quarterly*, 35, 557-582.
- Carroll, W. (1986) *Corporate Power and Canadian Capitalism*. Vancouver: University of British Columbia Press.
- Carroll, W. and M. Armstrong. (1999) "Finance Capital and Capitalist Class Integration in the 1990s: Networks of Interlocking Directorships in Canada and Australia", *Canadian Review of Sociology and Anthropology*, 36, 3, 331- 345.
- Carroll, W.K. and M. Fennema. (2002) "Is There a Transnational Business Community?", *International Sociology*, 17, 393-323.
- Chua, W.F. and R. Petty. (1999) "Mimicry, Director Interlocks, and the Interorganizational Diffusion of a Quality Strategy: A Note", *Journal of Management Accounting Research*, 11, 93-104.
- Coles, J.W. and W.S. Hesterly. (2000) "Independence of the Chairman and Board Composition: Firm Choices and Shareholder Value", *Journal of Management*, 26, 195-214.
- Committee on the Financial Aspects of Corporate Governance [CFCG] [Cadbury Report]. (1992) *The Financial Aspects of Corporate Governance*. London: CFCG.

- Canyon, M.J. (1994) "Corporate Governance Changes in UK Companies between 1988 and 1993", *Corporate Governance: An International Review*, 2, 97-109.
- Davis, G.F. (1991) "Agents Without Principles? The Spread of the Poison Pill Through the Intercompany Network", *Administrative Science Quarterly*, 36, 583-613.
- Domhoff, G.W. (1967) *Who Rules America?*. Englewood Cliffs, NJ.: Prentice-Hall.
- Dooley, P.C. (1969) "The Interlocking Directorate", *American Economic Review*, 59, 314-323.
- Fennema, M. and H. Schijf. (1979) "Analysing Interlocking Directorates: Theory and Method", *Social Networks*, 1, 297-332.
- Fligstein, N. (1985) "The Spread of the Multidivisional Form Among Large Firms, 1919-1979", *American Sociological Review*, 50, 377-391.
- Fligstein, N. and P. Brantley. (1992) "Bank Control, Owner Control, or Organizational Dynamics: Who Controls The Large Modern Corporation?", *American Journal of Sociology*, 98, 280-307.
- Freitag, P.J. (1975) "The Cabinet and Big Business: A Study of Interlocks", *Social Problems*, 23, 137-152.
- Galaskiewicz, J. and S. Wasserman. (1989) "Mimetic Processes within an Interorganizational Field: An Empirical Test", *Administrative Science Quarterly*, 34, 454-479.
- Geletkanycz, M.A. and D.C. Hambrick. (1997) "The External Ties of Top Executives: Implications for Strategic Choice and Performance", *Administrative Science Quarterly*, 42, 654-681.
- Haunschild, P.R. (1993) "Interorganizational Imitation: The Impact of Interlocks on Corporate Acquisition Activity", *Administrative Science Quarterly*, 38, 564-592.
- Haunschild, P.R. (1994) "How Much is That Company Worth?: Interorganizational Relationships, Uncertainty, and Acquisition Premiums", *Administrative Science Quarterly*, 39, 391-411.
- Haunschild, P.R. and C. M. Beckman. (1998) "When Do Interlocks Matter?: Alternate Sources of Information and Interlock Influence", *Administrative Science Quarterly*, 43, 815-844.
- Haveman, H.A. (1993) "Follow the Leader: Mimetic Isomorphism and Entry into New Markets", *Administrative Science Quarterly*, 38, 593-627.
- Koenig, T.; R. Gogel and J. Sonquist. (1979) "Models of the Significance of Interlocking Corporate Directorships", *American Journal of Economics and Sociology*, 38, 173-186.
- Kosnik, R. D. (1987) "Greenmail: A Study of Board Performance in Corporate Governance", *Administrative Science Quarterly*, 32, 163-185.
- Kotz, D. (1978) *Bank Control of Large Corporations in the United States*. Berkeley, CA.: University of California Press,
- Lang, J.J. and D.E. Lockhart. (1990) "Increased Environmental Uncertainty and Changes in Board Management Practices", *Academy of Management Journal*, 33, pp.106-28.
- Mace, M. (1971) *Directors: Myth and Reality*. Boston: Graduate School of Business Administration, Harvard University.
- Mariolis, P. (1975) "Interlocking Directorates and Control of Corporations: The Theory of Bank Control", *Social Science Quarterly*, 56, 425-439.
- McCarthy, D. and Puffer, S. (2002) "Corporate Governance in Russia: Towards a European, US or Russian Model", *European Management Journal*, 20, 630-640.

- McNulty, T. and A. Pettigrew. (1999) "Strategists on the Board", *Organization Studies*, 20, 47-74.
- Meyer, G.W. (1994) "Social Information Processing and Social Networks: A Test of Social Influence Mechanisms", *Human Relations*, 47, 1013-1048.
- Mintz, B. (1995) "Business Participation in Health Care Policy Reform: Factors Contributing to Collective Action Within the Business Community", *Social Problems*, 42, 408-428.
- Mintz, B. and M. Schwartz. (1981) "Interlocking Directorates and Interest Group Formation", *American Sociological Review*, 46, 851-869.
- Mintz, B. and M. Schwartz. (1985) *The Power Structure of American Business*. Chicago: University of Chicago Press.
- Mizruchi, M. (1982) *The American Corporate Network 1904-1974*. Beverly Hills, CA.: Sage.
- Mizruchi, M.S. (1992) *The Structure of Corporate Political Action: Interfirm Relations and their Consequences*. Cambridge, MA and London: Harvard University Press.
- Mizruchi, M.S. (1996) "What Do Interlocks Do? Analysis, Critique, and Assessment of Research on Interlocking Directorates", *Annual Review of Sociology*, 22, 271-298.
- Mizruchi, M.S. and L.B. Stearns. (1988) "A Longitudinal Study of the Formation of Interlocking Directorates", *Administrative Science Quarterly*, 39, 194-210.
- Mizruchi, M.S. and L.B. Stearns. (1994) "A Longitudinal Study of Borrowing by Large American Corporations", *Administrative Science Quarterly*, 39, 118-140.
- New York Stock Exchange. [NYSE] (2003). *Final NYSE Corporate Governance Rules*. New York: NYSE. www.nyse.com/pdfs/finalcorpgovrules.pdf
- O'Reilly, C.A.; B.G. Main and G. S. Crystal. (1988) "CEO Compensation as Tournament and Social Comparison: A Tale of Two Theories", *Administrative Science Quarterly*, 33, 257-274.
- Ornstein, M. (1984) "Interlocking Directorates in Canada: Intercorporate or Class Alliance?" *Administrative Science Quarterly*, 29, 210-231.
- Ornstein, M. (1989) "The Social Organization of the Canadian Capitalist Class in Comparative Perspective", *Canadian Review of Sociology and Anthropology*, 26, 151-177.
- Overbeek, H. (1990) *Global Capitalism and National Decline*. London: Unwin Hyman.
- Palmer, D.A. (1983) "Broken Ties: Interlocking Directorates and Intercorporate Coordination", *Administrative Science Quarterly*, 28, 40-55.
- Palmer, D.A.; P.D. Jennings and X. Zhou. (1993) "Late Adoption of the Multidivisional Form by Large U.S. Corporations: Institutional, Political, and Economic Accounts", *Administrative Science Quarterly*, 38, 100-131.
- Richardson, R.J. (1987) "Directorship Interlocks and Corporate Profitability", *Administrative Science Quarterly*, 32, 367-386.
- Scott, J. (1985) *Corporations, Class and Capitalism*. Second Edition. London: Hutchinson.
- Scott, J. (1991) "Networks of Corporate Power: A Comparative Assessment", *Annual Review of Sociology*, 17, 181-203.
- Sonquist, J. and T. Koenig. (1975) "Interlocking Directorships in the Top US Corporations", *Insurgent Sociologist*, 5, 196-230.
- Stearns, L. and M.S. Mizruchi. (1986) "Broken-tie Reconstitution and the Functions of Interorganizational Networks: A Reexamination",

- Administrative Science Quarterly*, 31, 522-538.
- Stockman, F.N.; R. Ziegler and J. Scott. (1985) (Editors) *Networks of Corporate Power: A Comparative Analysis of Ten Countries*. Cambridge, UK: Polity Press.
- Sweezy, P.M. (1953) "The Decline of the Investment Banker", in P. M. Sweezy, *The Present as History: Essays and Reviews on Capitalism and Socialism*. New York and London: Monthly Review Press, 189-196.
- United States House of Representatives Subcommittee of the Committee on Banking and Currency [Pujo Commission]. (1913) *Investigation of Financial and Monetary Conditions in the United States*. Washington, DC.: US Government Printing Office.
- Useem, M. (1984) *The Inner Circle: Large Corporations and the Rise of Business Political Activity in the US and UK*. New York: Oxford University Press.
- Wade, J.; C.A. O'Reilly and I. Chandratat. (1990) "Golden Parachutes: CEOs and the Exercise of Social Influence", *Administrative Science Quarterly*, 35, 587-603.
- Westphal, James D.; Ranjay Gulati and Stephen M. Shortell. (1997) "Customization or Conformity? An Institutional and Network Perspective on the Content and Consequences of TQM Adoption", *Administrative Science Quarterly*, 42, 366-394.
- Westphal, James D.; M.-D.L Seidel and K.J. Stewart. (2001) "Second-order Imitation: Uncovering Latent Effects of Board Network Ties", *Administrative Science Quarterly*, 46, 717-749.
- Westphal, James D. and E.J. Zajac. (1997) "Defections from the Inner Circle: Social Exchange, Reciprocity, and the Diffusion of Board Independence in U.S. Corporations", *Administrative Science Quarterly*, 42, 161-183.
- Windlof, P. (2002) *Corporate Networks in Europe and the United States*. Oxford: Oxford University Press.
- Zeitlin, M. (1974) "Corporate Ownership and Control: the Large Corporation and the Capitalist Class", *American Journal of Sociology*, 79, 1073-1119.

Bruce Cronin

*Department of International Business & Economics, University of Greenwich
London, UK
B.Cronin@gre.ac.uk*

Justice, Morality and Ethics

John Davis

Introduction

The concepts of justice, morality, and ethics provide both a normative framework for national and international public policy and at the same time important means by which public policy is interpreted in the governance structures of government, private business firms, and non-governmental organizations. Because public policy is necessarily formulated in broad terms meant to be comprehensive of a variety of different circumstances, it typically leaves its manner of application incompletely specified. The concepts of justice, morality, and ethics, as a framework in which public policy is formulated, then constitute a key resource and guide in the making of concrete governance decisions with respect to interpreting the intent of policy-makers. Different societies and different groups within societies, of course, have different normative commitments, and this might be thought a barrier to these normative commitments playing a role in the interpretation of public policy. One response to this is to say that even conflicting normative commitments provide resources for governance, because disagreements are still defined in shared vocabulary and language ('this is fair to do' vis-à-vis 'this is not fair to do') which presuppose shared normative concerns. A second view is to say that governance should be seen as a contested process in which peoples' different normative commitments compete to determine the way in which public policy is ultimately implemented.

However the governance process is understood, in its dependence on these particular normative concepts and commitments, it needs to be seen as informal in nature rather than rule-governed type of

deliberation in the sense that individuals typically rely on their ordinary intuitions regarding what they believe to be just, moral, or ethical rather than make use of systematic writings or literatures on these subjects when they seek to interpret and implement public policy. Thus, a debate over, say, whether a particular practice constitutes insider trading, might couch technical-legal questions regarding whether the directors and managers of a firm were in violation of a country's laws on insider trading in a set of judgments about whether or not their actions might be thought 'fair' to outsiders. The question of 'fairness' might then be further evaluated at different levels, such as whether expectations of equal access to information or exercise of property rights in information were abridged (Koslowski 1995). Though case studies in business schools and philosophy courses may map out the complex interconnections between these different and often overlapping dimensions and levels and individuals' normative commitments after the fact, the social process of governance in the world of business and government is more practically-oriented, and must accordingly draw on the general understanding of values of those immediately involved. Thus justice, morality, and ethics in the framework of governance questions need to be looked at more pragmatically as all-purpose tools to be used when thought likely to contribute solutions and as needed.

It is possible to see better how this comes about by attending to the scope within which each of these fundamental normative concepts finds its most standard application. This will in turn make it possible to classify distinct types of governance issues according to differences in the way these normative concepts are employed. In reverse-order of the concepts as listed above, then, the scope of ethics generally concerns individual decision-making, the scope of morality

concerns the character of groups, institutions, and socioeconomic processes involving many people, and the scope of justice concerns the evaluation of entire socioeconomic systems. The scope of the three concepts, that is, ranges from the micro level for questions of ethics to an intermediate or meso level for matters of morality to the macro level for issues of justice. Correspondingly, governance questions can be distinguished according to whether they apply to the ethics of individual decision-making, to the morality of groups, institutions, and socioeconomic processes, or to entire socioeconomic systems in the global political economy. In each instance the location at which governance issues arise also determines which people those issues concern as well as the nature of their involvement in them. The three cases are distinguished with examples in the next three sections.

Following this, these normative concepts and their informal roles in governance processes are compared to a more formal and systematic application of different sort of normative reasoning in connection with national government economic policy-making in the form of efficiency and cost-benefit recommendations made primarily by professional economists, especially since World War II. Here highly determinate methods are employed by experts in public economics and public finance to generate specific policy recommendations which are grounded in well-defined normative criteria. Because professional expertise dominates this type of policy recommendation, the governance process is narrower in terms of who it involves than is the case when individuals in many locations and different positions in a society draw informally on their ordinary intuitions regarding what they believe to be just, moral, or ethical. The combination of this more informal normative reasoning with the more professionalized type

of policy recommendation associated with efficiency recommendations gives countries a two-level type of social governance process that has implications for the nature of democracy in those countries. Across countries, or in the global political economy, where there exists different kinds of political institutions, the combination of these two levels of social governance takes on a somewhat different shape. Both national and international cases are discussed and distinguished below. The article concludes with brief comments on international economic integration.

Ethics

The subject of ethics includes such concepts as obligation, right, wrong, and good, and generally concerns how individuals decide what they morally ought to do (MacIntyre, 1966). One example of the role of ethics in individual decision-making as a guide to governance that applies to business and organizational ethics is the question of whistle-blowing (Jensen, 1987). Whistle-blowing involves individuals making public firm or organization practices which are either in violation of law or seen to be socially unacceptable for any number of reasons. Whistle-blowers are individuals with inside knowledge of such practices who elect to publicize them at personal risk to themselves. Ethics arises in two ways in connection with whistle-blowing. First, individuals who find themselves in whistle-blowing situations must decide whether to take action when it may not be in their own interest. Thus they must decide whether to place what they regard as the right thing to do in balance against self-interest. Second, firms and organizations must decide whether they will foster environments which encourage or discourage the actions of individuals as whistle-blowers. They thus not only also balance self-interest with the question of what

the issue of right and wrong, but also bear responsibility for easing or alleviating the decisions of individuals faced with their own ethical choices. That is, if the environment in a firm or organization is hostile to whistle-blowing, it can be argued that the situation is doubly wrong both on account of the practices that escape public view and on account of the additional burdens placed on individuals concerned with their responsibilities as possible whistle-blowers.

A second example of ethics that applies to households concerns individuals' end-of-life decision-making for family members (Rosenfeld *et al.*, 2000). As medical technologies increasingly make it possible to sustain life, even when the chances for return to normal living are small, individuals find themselves more often faced with questions regarding how much responsibility they should take for others' lives. Health professionals are often constrained legally to make every effort to preserve life, but family members may find themselves in the position of needing to act on a loved one's prior expressed intentions when those individuals are incapacitated. In many instances, health power of attorney and living wills are agreed upon in advance giving responsibility for health decisions to a designated family member. In these cases, a legal agreement sets guidelines for care, and a governance structure exists that helps define individual choices. But general guidelines may not offer enough direction when there is significant medical uncertainty. In these cases, individuals who bear responsibility for decision-making must ask themselves what the right thing is to do, where this involves trying to understand what an incapacitated family member would have wished be done. An informal type of governance may be of some assistance. The decision-making individual can rely on standards and norms in society generally as operates through

counseling and discussion with others. But cases such as these remain highly ethical in that the general character of such direction ultimately leaves some individual with the responsibility to decide what is right or wrong to do.

Note that public policy in both cases may also be hard to define. In the first case, if firms and organizations are not actually in violation of the law, or if it is unclear whether this is the case, it can be argued that the default solution for the potential whistle-blower is to do nothing. An individual, however, may believe that the spirit of the law or the more general thrust of public policy calls for action. For example, a broad consensus exists in many societies that protecting the environment is a high priority. Since damage to the environment is often hard to predict, an individual may believe it ethically wrong to not bring simply the risk of damage to wider attention. In the second case, individuals responsible for the care of others usually rely on the opinions of medical professionals. At the same time, medical professionals' advice does not extend to ethical issues surrounding the consequences of care. More generally, attitudes toward health and life vary across society. Thus family members occupy a unique position in this domain of ethical decision-making.

Note also that no sophisticated ethical reasoning is generally required for these types of decisions. An individual may rely on ordinary intuitions about the social good or responsibility to others that derive from common experience. Finally, note that the micro character of governance in cases such as these is reflected in the focus on individual judgment where there limited guidance from others.

Morality

Morality concerns normative relationships that operate within social groups, institutions,

and socioeconomic processes involving many people. One example is the issue of child labor in the contemporary global political economy. Child labor is an important problem in many societies, but not systematically reflective of entire socioeconomic systems. Thus most contemporary societies regard child labor as a morally unacceptable state of affairs, and look critically upon firms and social arrangements that make use of it or facilitate it. This value is rooted in a broad normative conception of human rights and human dignity, as well as in a general understanding of what is thought to be necessary for people to realize their potential as human beings. But while questions of ethics arise with respect to what individuals might do to prevent child labor, the primary focus in this case involves identifying where and under what circumstances child labor is practiced, and in then bringing moral standards to bear in an effort to eliminate it. In short, child labor is a practice or institution which needs to be seen as a form of social interaction at the meso level.

Public policy regarding child labor in the world today operates on different levels and with different powers of enforcement. Most nations have adopted legislation against child labor, but the degree of protection they afford can be quite uneven (Lieten 2001). At the same time, international covenants, such as the 1989 United Nations Convention on the Rights of the Child, offer broad guidelines regarding exploitation and harmful work without specifying how these concepts are to be measured. The consequence of this is that nations with low standards can claim that they protect children according to their own institutions and laws though their levels of protection can be found deficient relative to the higher standards of other nations. This creates an occasion for public governance in that the widely shared ambition of eliminating

child labor in principle can be articulated in terms of an ascending set of goals that can be increasingly implemented over time across countries. That is, ambiguity internationally over the force and appropriate application of public policy in regard to child labor shifts responsibility to societies to further define how they plan to address the problem of child labor over time. That people from many different social values systems across the world might participate in such a process implies that the normative principles brought to bear are general ones that can be widely shared. Morality as a framework for public policy thus underlies and supports a social governance process which in turn helps further define and sharpen public policy.

A second example where morality influences behavior is the issue of excessive executive compensation (Murphy 1999). Executive compensation is argued to be excessive when its level cannot be readily explained by the productivity of the individuals involved, such as when compensation increases though firm performance deteriorates. The issue is one that operates at a meso level and does not apply to entire socioeconomic systems, since the ratio of executive compensation to average wages has grown faster in some firms, industries, and economies than others. Further, since executive compensation is measured relative to average wages, it is an issue that concerns groups of people and a particular kind of institutions.

The normative value operating in connection with executive pay is also one of rights, though in this instance in the form of the rights of shareholders. Shareholder rights have different interpretations and vary in scope across countries. Thus public policy on the subject of excessive executive compensation is formulated in ways that take into account these differences. In societies where shareholder rights are narrowly defined

on this subject, as in the United States, the issue is often addressed on a firm-by-firm basis, such as when institutional investors express concern regarding levels and methods of executive compensation. In societies in contrast where a wider scope is given shareholder rights, as in the United Kingdom, Australia, and the Netherlands, shareholder activism is embedded in laws permitting non-binding resolutions on executive pay. Different countries' social values thus help define different approaches to public governance. In contrast to ethical reasoning, then, morality is less a matter of individual decision-making and more a matter of how groups of people approach a matter.

Justice

Justice or fairness is often thought to be the most important normative concept in contemporary society. Indeed as a concept that provides normative understanding of entire socioeconomic systems, its reach and significance cannot be denied. At the same time, there are a variety of theories regarding what justice and fairness mean, so that not surprisingly philosophical scholarly debate about the concept is extensive (Lamont and Favor 2007). However, two very central and commonly held understandings of justice are employed to illustrate their contribution to creating a normative framework for public policy and guide to social governance, namely, justice is understood as reciprocity (in popular terms, the Golden Rule) and in terms of equality.

Reciprocity as a principle exists in most societies, and it has been argued by some biologists that it is not simply a result of learned behavior (Trivers 1971). In recent years, it has become a subject of intense experimental investigation in economics in regard to its pervasiveness in human thinking (Fehr & Gächter 2000). Whereas economics has traditionally seen self-interest as the

dominant motivation in economic action, experimental research strongly suggests that people will put aside self-interest to respond to others in the manner that they themselves have been treated, whether for ill or for good.

In most nations, public policy also enshrines principles of justice as reciprocity in the most basic way in the form of constitutional legal foundations. While expert authorities in government, business, and law have primary responsibility for interpreting these foundations, they do so in an environment in which most people have views about just and fair behavior. This gives social governance where justice is concerned the combined character of being expert and popular at the same time. For example, consider the history of discrimination law in the United States, particularly as applies to women in the labor force (Bergman 1986). Changes in social values and family structure were an impetus to higher rates of labor force participation for women; business firms, accustomed to work forces dominated by men, found themselves charged with unfair treatment of women; legal challenges regarding employment practices became more common, and the courts were brought in; eventually laws were changed that clarified and changed the legal environment. Thus new standards regarding fair and just treatment of women were the product of a governance process that pervaded society and its institutions. Public policy seen in this light is a distillation of a complex of forces linking governance and principles of justice.

Equality as a normative principle is the idea that like individuals should be treated in a like manner. As a principle of justice it is also deeply embedded in many countries' legal and political systems, and thus is also relevant to entire socioeconomic systems. For example, consider the issue of equality of educational opportunity (Coleman 1990), which is associated with many countries'

historical economic development. When the opportunity for education is restricted to only a small part of a country's population, economic development is impeded. In the process of development, however, universal education becomes an ideal, and is often defended as a matter of justice. Social governance, then, particularly as it applies to the allocation of resources, draws on normative principles that are seen as relevant to entire societies. Like the reciprocity understanding of justice, the equality one is subject to many interpretations. But this gives these principles a versatility that allows many to adopt them in debate over social reform.

Efficiency and Cost-Benefit Judgements

Efficiency recommendations are formulated in terms of the Pareto optimality criterion which states that any proposed re-allocation of resources is recommended when it makes at least one person better off in the sense of being strictly preferred by that person without at the same time making someone else worse off. One attraction of the Pareto criterion for economists is that it appears uncontroversial in nature in simply favoring unequivocal improvements in well being. A related attraction, then, is that as the Pareto principle appears to be independent of these other sorts of normative principles, it is relatively straightforward to apply. Economists. This leads economists with their specialized knowledge in the theory of markets and mastery of technical research to treat efficiency recommendations as a domain of expert analysis.

Cost-benefit analysis and the recommendations based upon it share with Pareto efficiency recommendations a focus on enhancing well-being. Like the Pareto criterion, cost-benefit evaluations appear essentially pragmatic in nature and free of significant normative content. At the same time, cost-benefit evaluations, while

straightforward in principle, also depend on expert knowledge associated with the problem of fully capturing all costs and benefits in money terms. Economists who are practiced in public economics have developed a variety of conventions and rules for carrying out analysis of this sort, and as in the case of Pareto recommendations this tends to limit access to this part of the policy-making process to them while excluding others without the necessary technical expertise.

That the governance process in the case of Pareto recommendations and cost-benefits analysis is socially narrower than is the case when individuals draw informally on their ordinary intuitions regarding what they believe to be just, moral, or ethical is thus due to the role of expert knowledge. Further, that expert knowledge has this role is due to the assumption that Pareto recommendations and cost-benefits analysis are essentially pragmatic in character and accordingly free of significant normative dimensions. Given the nature of ethics, morality, and justice as broad principles widely employed, it can be said that overall the social governance process operates on two different levels: one that draws specifically on the specialized knowledge of experts, and another that relies on general understanding of general normative principles.

Social Governance in the Global Political Economy

Across countries where there exist different kinds of political institutions, the expert and more informal domains for public policy with their associated normative commitments combine somewhat differently. On the one hand, the status of economics as shared theory of the global political economy is high, particularly because standard economics emphasizes the market process, and globalization is widely seen as the elimination of barriers to markets as well as the extension

of markets into incompletely monetized economies. This promotes economists' emphasis on efficiency and well-being gains as a principal objective of economic development. At the same time, since political institutions differ across countries, shared views regarding other normative principles are comparatively weak. Not only do different societies with their different cultures exhibit different mixes of normative principles, but their interpretation of these other principles can vary in significant ways. Thus social governance in the global political economy tends to rely more heavily on efficiency recommendations and economists' expert advice than is the case in individual countries. Indeed, though Pareto recommendations are not by themselves sufficient for determining public policy in either realm, their perceived appearance as pragmatic and uncontroversial can lend them a default status as definitive guides for international social governance in the face of often deep differences between countries regarding other normative principles.

On the other hand, though the logic of Pareto efficiency recommendations is widely accepted in the global political economy, there are nonetheless important limits to its application associated with the political organization of the world in terms of separate nation states. This is particularly evident with respect to the application of Pareto reasoning to the problem of negative externalities in markets. Negative externalities exist when market transactions have costly consequences for third parties – an important example of which is environmental pollution. Economists use Pareto reasoning to argue that pollution should be eliminated by demonstrating that well-being can be increased by 'internalizing' the costs of the externality or having the parties to the market transaction responsible for that externality absorb its costs. However, many externality problems in the world

economy have extra-territorial dimensions in that third party effects occur in countries different than the countries in which the transactions occurred that were responsible for those subsequent effects. Examples of these 'super-externalities' (Dua and Esty 1997) range from greenhouse gas generation by individual countries that affects the entire world to more local spillovers such as one country's over-fishing practices that depletes maritime resources shared by a number of countries. In cases such as these, Pareto recommendations must be complemented by further policy principles which can be employed to arbitrate between competing national interests. Historically, these further principles have been understood in terms of nations having certain legal rights in the international community associated with participation in international institutions such as the United Nations and World Court. Though the force of such rights is often tenuous, they nonetheless play a role complementary to Pareto recommendations.

In the global political economy, then, economics as shared theory of markets is integrated with the political organization of the world in terms of nation states. While in individual countries economists' Pareto reasoning has advantages over more informal normative reasoning with respect to such matters as justice, morality, and ethics, in the global political economy national interest constitutes a more considerable counterweight to economists' emphasis on efficiency. An important question, then, is how the future development of the global political economy might affect both the international balance between these principles and also that within individual countries. The issue has been addressed by Dani Rodrik (2000) in his projection of three possible futures for the global political economy that differently combine two but not three of the

following: the nation state, economic integration, and democracy.

First, the future may include continuation of nation states as sovereign powers in the world and economic integration at the expense of democracy within countries (termed the 'Golden straitjacket'). Essentially Rodrik argues that nations could compete with one another for economic resources in the global political economy, but that this would be incompatible with historical democratic traditions in many of them. In terms of the balance between Pareto efficiency recommendations and the more informal normative reasoning associated with justice, morality and ethics, this would mean further reliance on the former as economists as expert authorities would be called upon to explain the requirements of integrating national economies into the global political economy. Second, the future may preserve the nation state and historical democratic traditions, such that the current combination of Pareto thinking and more informal normative principles in the social governance process continues at the expense of further economic integration between nations (termed the 'Bretton Woods compromise'). In effect, the status quo largely prevails, or at least nations are not drawn into further economic linkages. Third, nation states cease to be important agents in the future, economic integration proceeds ahead, democratic traditions also continue, but are re-formed across and within countries (termed 'Global federalism'). In effect, the current balance within nations between Pareto thinking and the more informal normative reasoning associated with principles of justice, morality and ethics is re-constituted across nations.

Of course whether one of these futures will come about cannot be determined at present. But Rodrik's possibilities are instructive for considering how the principles of justice, morality, and ethics may be integrated with

efficiency recommendations in public policy at both national and international levels. One lesson that is clear from this is that how all these principles combine reflects the institutional characteristics of countries and the global political economy. The social governance process, whether in its national or international form, employs normative principles that reflect historical developments that have come to be embedded in institutional forms of political and economic organization. Justice, morality, and ethics, then, though often thought primarily abstract investigation, ought as much be seen as practical principles of everyday concern that reflect dilemmas and conflicts of the social governance process.

Selected References

- Bergman, Barbara. (1986) *The Economic Emergence of Women*. New York, Basic Books.
- Coleman, James S. (1990) *Equality and Achievement in Education*. Boulder, CO: Westview.
- Dua, André and Daniel Esty. (1997) *Sustaining the Asia Pacific Miracle, Environmental Protection and Economic Integration*. Washington, DC, Institute for International Economics.
- Fehr, Ernst and Simon Gächter. (2000) "Fairness and Retaliation, The Economics of Reciprocity", *Journal of Economic Perspectives*, 14, 159-181.
- Jensen, J. Vernon (1987) "Ethical tension Points in Whistleblowing", *Journal of Business Ethics*, 6, 4, 321-328.
- Koslowski, Peter (1995) "The Ethics of Banking, On the Ethical Economy of the Credit and Capital Market, of Speculation and Insider Trading in the German Experience", in Argandona, Antonio (Editor), *The Ethical Dimension of Financial Institutions and Markets*. Berlin, Springer, 180-232.

- Lamont, Julian and Christi Favor. (2007) “Distributive Justice”, in Edward N. Zalta (Editor), *Stanford Encyclopedia of Philosophy*. plato.stanford.edu
- Lieten, G. Kristoffel. (2001) “Child Labour. Questions on Magnitude”, in G.K. Lieten and Ben White (Editors), *Child Labour. Policy Perspectives*. Amsterdam, Aksant Academic Publishers, 49-66.
- MacIntyre, Alisdair. (1966) *A Short History of Ethics*. London, Routledge Kegan Paul.
- Murphy, Kevin J. (1999) “Executive Compensation”, in O. Ashenfelter and D. Card (Editors), *Handbook of Labor Economics*. Volume 3b. Amsterdam, Elsevier Science North Holland, 2485-2563.
- Rodrik, Dani. (2000) “How Far Will International Economic Integration Go?”, *Journal of Economic Perspectives*, 14, 1, 177-186.
- Rosenfeld, Kenneth E.; Neil S. Wenger and Marjorie Kagawa-Singer. (2000) “End-Of-Life Decision Making. A Qualitative Study of Elderly Individuals”, *Journal of General Internal Medicine*, 15, 9, 620-625.
- Trivers, Robert. (1971) “The Evolution of Reciprocal Altruism”, *Quarterly Review of Biology*, 46, 35-56.

John B. Davis
Marquette University
Milwaukee, USA;
University of Amsterdam
The Netherlands
john.davis@mu.edu
j.b.davis.uva.nl

Land-Use Governance

*Klaus Hubacek, Evan Fraser
and Shova Thapa*

Introduction: Multidimensional Land

The phrase “land use governance” seems to imply that land is a single monolithic entity and that all people who use land can be governed by roughly the same set of norms and standards. This is not the case. Land is an aggregate of many different attributes, providing many important functions, many of which are not part of market transactions. For example, the same piece of North American prairie may have value as farmland, contain natural gas or other mineral deposits, be part of an urban drinking water supply, be wildlife habitat, and have aesthetic values. An analysis of land use governance has to include the unique character of land comprising its biophysical, socio-economic and institutional properties.

Land provides economic goods or source functions (Thomas et al 2000) such as agriculture, fisheries and mineral ore. It can provide economic benefits since certain types of pollution are assimilated by land (e.g. carbon dioxide is sequestered into plant material). Since land is the medium through which energy and nutrients cycle, land also provides a host of ecosystem functions and this affects the resilience of ecosystems (defined as the ability to recover specific function conditions after a disturbance).

It is also possible to view land as being comprised of various forms of capital that include varying degrees of pure natural capital but also human-made capital, which results from previous investments in land reclamation, drainage, and soil improvements. The human capital component is of increasing importance in developed countries where increasing amounts of investments are made to prepare land for

economic activity as land takes on increasing levels of value. In the developing world human activities and livelihoods are closely linked with the natural environment. In these situations, communities directly depend on the continued productivity of natural resources such as timber and wildlife often enhanced by capital investment. The “capital” view of land also relates to land use, which refers to those human activities that exploit the various resources provided by the (biophysical) features of an area. Meyer (1999), however, cautions that land use is such a broad term (after all, all human activities depend on land to a greater or lesser extent) that it could become unmanageable and meaningless.

An alternative way of understanding land and land use is in terms of the institutional arrangements that govern the rights and responsibilities different groups have over specific pieces of land. For example, the existence of parcels of land (usually referred to as “real estate”) is a matter of human institutions, and comes into existence as a result of historical, economic and social factors that set the “rules of the game.” Public regulations, such as community plans, zoning ordinances, rent controls, subdivision regulations, building codes, and laws pertaining, for example, to mortgage finance, shape the development and use of land. Contributing to this institutional setting are cultural, economic, political, religious, social, and traditional factors. Less tangible institutions are customs and traditions, which are the way of thinking and acting within a certain religion and culture (Hubacek and van den Bergh 2002; Hubacek and Vazquez 2002).

One way of untangling these various and sometimes competing definitions is to separate the various goods that land provides and characterize each of these goods as “subtractable” and “excludable”. Subtractable

goods are those that are used up by a single user (minerals are used once extracted, while amenity value is not used up by being viewed). Excludable goods are those that one user can easily exclude other potential users from (for example, it is relatively easy to fence off farm land, while ground water that flows freely through watersheds is much more difficult to exclude others from) (Ostrom et al 1999; Ostrom et al 1994). Different strategies will be required to promote these different types of goods. To preserve goods that are both subtractable and excludable (such as soil fertility that is used to produce agricultural commodities) land ownership is necessary so that a farmer who invests in long-term management practices will receive the benefits of this investment (Lee 1980; Lumley 1997; Praneetvatakul et al 2001). Long-term tenure, however, does not provide incentives for farmers to preserve the non-excludable and non-subtractable “public goods” their land produces such as wildlife habitat (Fraser 2003). In this case, other incentives must be devised.

Whichever of these lenses one chooses to apply to land, it is clear that serious problems exist in the way human activity treats this resource. Society has changed and modified the structure, composition, attributes and functions of land around the world. The conversion, fragmentation, degradation and exploitation of land today are the source of many threats to the global environment such as biodiversity loss, species extinction, global warming and climate change. This is partly because we depend on land for so many things that our demands sometimes conflict and decisions are often based on short-term economic gain. Part of these problems emerge because our present political economic system treats land a commodity to be bought and sold to the highest bidder and to be used at its highest economic values for the landowner. Our struggle to capture

increasing economic values leads to us forgetting that land has many (often times non-economic) uses and is a finite resource.

Land as Commodity: Market Transactions

Since market transactions are clearly one of the most dominant ways to govern land use, it is helpful to unpack some economic issues. Within the logic of the market economy, land has been reduced to a factor of production and an object of consumption. Different functions of land and land resources do not have any value per se and are only valued as revealed by final demand. The physical qualities of land are reduced in economics to the willingness-to-pay of economic agents represented in market transactions. Private production and consumption decisions, such as the allocation of land or resources between alternative uses, are made with the objective of maximizing utility accruing to individual producers or consumers, subject to constraints imposed by prevailing technology, resources, and policies. According to this logic, land-use decisions are mainly governed by supply and demand via the price mechanism.

The economic supply of land depends itself on a number of factors: physical supply, institutional arrangements, available technology, and location. Economic supply may be defined as land units that enter particular uses in response to certain stimuli, such as prices and institutions. The owner of land decides the type and intensity of use dependent on the price the land will bring on the market. The present economic supply reflects current utilization practices, current economic availability, and current adaptability of the material base to required demand.

Supply problems do not develop as long as each type of use can expand. Complications only arise when conflicting uses compete for the same land areas. Whereas supply is to some extent fixed, demand seems to be

unlimited. The fulfilment of all our needs is based on land. Each demand category has direct and indirect land requirements for the production of all the inputs necessary to produce the final product. Following this logic, the demand for land could be divided into two different categories: direct demand and derived demand. Direct demand for land is the demand for land that is used directly for consumption of land, guided by market signals such as land prices and land rent regulating supply and demand on real estate markets. Derived demand for land comes through the implicit market signals on good and factor markets that consumers give to land users, such as farmers, as to what land uses will satisfy current demand for goods and services.

The amount of land producers need to sustain the production of goods is directly influenced by the signals they receive from their customers by the way of prices. For instance, land resources tend to gravitate to those uses that command the highest market prices and offer the highest net returns to investment. Rising price levels usually encourage bringing more land into use and using the land already in use more intensively.

As mentioned above, producers treat land as an input to production. Land operators will try to find the proportion of inputs that derive satisfying returns for them. Therefore, they will evaluate the marginal value of their land not only by itself, but also in comparison with the marginal value of other factor inputs. This necessitates that substitution between factors of production is possible. In the short-run, producers are unable to make this substitution between land and other factor inputs because land is a relatively fixed factor of production. This stems from the fixed location of real estate resources, the ownership rights applied to them, and other institutional factors that make changes of production sites difficult.

From an investor's point of view, the rent paid by the user of real estate compensates for the investors opportunity costs (defined as the returns that could be received from alternative investments). Land rent represents the economic return that accrues to land for its use in production. Differences in rent-paying capacity or different classes of land are often explained in terms of different locations (e.g. closeness to water, infrastructure, amenities, and cultural centres) or different qualities of land (e.g. soil types or climatic factors and human-made improvements, such as buildings).

Land resources are at their highest and best use when they are used in a manner that provides an optimum return to their operators or to society. The highest and best use is subject to change in the quality of the land resource, changes in technology, changes in the demand structure, or changes in zoning ordinances or other legal institutions. In modern societies, land resources usually earn a higher return when used for commercial or industrial purposes than for any other purpose. The more highly valued and more economically productive uses usually take the better lands, leaving the lower-priority areas for other uses. Continuing expansion of high-priority lands leads to a discrimination of the economic supply of land available for other users and reduces idle land for undisturbed succession of the environment.

Externalities

While economic analysis typically assumes that land operators bear all the costs of and benefits from their activities, individual action usually affects third parties. For instance, if developers do not consider the loss of welfare to other people caused by their project, an external effect exists. An externality (a negative externality in this case) exists every time the action of one individual negatively affects the welfare of another and

the latter is not compensated for these losses. Externalities are very frequent in land use given the multiple products and costs often associated with uses of land resources. For instance, forests can be used for timber production, recreation, watershed protection, and wilderness, and often it is not possible or is too costly to avoid interference in these different uses. This has to do with the public good character of many environmental goods. The main feature of these goods is open access, that is, nobody can be excluded as soon as a good is provided, which leads to over-consumption.

Public goods such as land are often under-supplied by the market, thus requiring government intervention. Government can exercise its power to influence land-use decisions in many different ways: taxation, investment and subsidies, public ownership, and most importantly zoning. The idea of zoning is dividing land into districts having different land use regulations. Thus, many of the negative effects of physical interdependencies in production and consumption can be reduced by keeping sensitive and interacting production and consumption activities spatially apart. Other important land-use controls are subdivision controls, which impose restrictions to developers of land, and building and housing codes, which regulate construction, maintenance, and use of structures.

Text Box. Economic Incentive Measures and Biodiversity Conservation: Example of the Canadian Water Fowl

Land tenure is often cited as a way of ensuring sustainable land management. However, results of a study that compared land management practices on rented versus owned fields illustrated that while land tenure is necessary to sustainably produce some goods, it does not provide sufficient

incentives for all the different types of goods that land can produce. In South Western British Columbia, Canada, farmland produces both private goods, which are subtractable and excludable, and public goods, which are neither subtractable nor excludable. Subtractable goods include the agricultural commodities that a farmer sells to obtain an income, while public goods include wildlife habitat. Both are extremely important in this region that is one of the most productive region in Canada and sits on a major migration route for water fowl (Canada 1992). A recent study examine whether farmers with different types of land tenure were producing both types of good. Results of this study show that land ownership is necessary before farmers will plant crops that protect long-term soil fertility: farmers who rented their land planted considerably more annual cash crops that create long-term soil conservation problems compared with farmers who owned their fields. Owner-operators planted more soil-enhancing forage legumes and perennials. Land tenure, however, did not affect whether farmers encouraged bird habitat. Indeed, many farmers view local bird life as a liability since migrating flocks of ducks damage crops. To encourage farmers to promote this public good a separate policy was devised, and farmers are now paid a small amount per hectare to plant grasslands and winter forage crops that provide habitat for migrating ducks, swans, and geese (Delta Farmland & Wildlife Trust 2000, 2001).

A different possibility for dealing with conflicting land-use options is negotiation between stakeholders. In this process, all of the groups and individuals who are potentially affected by a land development project are invited to discuss the implications, alternatives, and modes of compensation.

Rights and Responsibilities of Different Types of Land Users

Although economic analysis usually assumes that the operators of land have unlimited freedom as to what resources they use and how and when they use them, in reality this freedom is restricted by the nature of their rights to use the land. These property rights refer to a bundle of entitlements defining the owner's rights, privileges, and limitations in the use of a particular land-related resource. Therefore, property rights or ownership is by far the most powerful institutional constraint guiding the operation of land in most contemporary economic systems.

Just as any given piece of land will provide different types of goods, there are also different types of users, each whom have different rights and responsibilities. At the most basic level, there are people who simply have the *authorized access* to a property. This type of user, who may use an area for recreation, has no claim over any subtractable or excludable goods the area produces and generally has the responsibility to ensure no damage is done to the region due to their access (e.g. hikers must not litter or remove plant/animal produce from national parks). There are also users who are authorized both to access a region and to extract certain goods. This would include people who have fishing, hunting, and timber licences. These *authorized users* would not typically be allowed to determine the quantity of the resource that they are entitled to and would have a responsibility to harvest in accordance with rules and codes of practice. The third type of user is called a "*claimant*" who has the right make decisions determining the rules governing the management, transformation or improvement of resources. *Proprietors* are the fourth type of user and have same rights as claimants as well as right to determine who may harvest the resources and who may be a claimant (this is called the right of exclusion).

The final type of user is the *owner* who has all of the above rights (table 1) as well as the right to transfer ownership rights to other people (i.e. selling a resource to another owner). This is called the right of alienation (Ostrom 2001).

Table 1 illustrates the different types of user of a resource (horizontal axis) and the different types of rights each user would have (vertical axis) taken from (Ostrom 2001).

Table 1. Users and Rights vis-à-vis Resources

	Owner	Proprietor	Claimant	Authorized User	Authorized Entrant
Access	X	X	X	X	X
With-drawal	X	X	X	X	
Management	X	X	X		
Exclusion	X	X			
Alienation	X				

Text Box. Land Use as a Bundle of Rights: Example of Urban Agriculture in Bangkok

An urban agriculture project in Bangkok, Thailand illustrated the need for policy to include an understanding of different rights and different users. In an effort to reduce urban poverty, a coalition of environmental Non-Governmental Organizations (NGOs) from Thailand and Canada obtained funding from the Canadian International Development Agency to establish urban agriculture projects with low-income communities in Bangkok, Thailand. To do this the International Centre for Sustainable Cities and the Thailand Environment Institute obtained permission from a group of landowners for low-income residents to access under-utilized land along the bank of a canal in the north end of Bangkok. This land was cleared of debris and existing vegetation and planted to a number of small vegetable plots that were initially used by individual households for subsistence. After a while, the community

decided to set up a roadside concession and sell the produce of the vegetable gardens. The landowner objected to this and stopped allowing the community to have access to the land. This situation can be explained using the typology introduced above. After the initial request, the landowners allowed the communities the right of access to their land, making the community members authorized entrants. The landowners did not, however, give the communities rights of withdrawal or of management, so when the community members decided to change what they were using the land for, and sell the produce of the gardens, the landlord decided that the community members were claiming more rights than they were entitled to. As a result, the owners then used their rights of exclusion, and barred the communities from even entering the area (Fraser 2002). This problem may have been averted if at the outset the NGOs had negotiated rights of withdrawal and management for the communities.

Once we accept that any given area of land may have different user groups, it is necessary to develop ways of resolving conflicts between them. The primary method we have for this is to develop land use plans using a broad base of stakeholder consultation and dialogue. This philosophy is enshrined in the United Nations' *Local Agenda 21* (United Nations 1992), which clearly states that all people have to play a role not just in the 'doing' of a development project but in setting the agenda as to what development means. Stakeholder consultation means that planners and those who establish the norms and standards that govern land use issues, work with a large number of constituents to establish targets, defined goals, and use this information to create plans to move towards collective visions (Coast Information Team 2002).

Text box. Land Use Conflicts and Stakeholder Involvement: Establishment of the Trans-Boundary National Park Neusiedler See – Seewinkel

The history of the now National Park has been a history of land use conflicts between landowners and multiple use such as fishing, reed grass extraction, hunting, tourism, residential areas, agricultural use. First steps to protect the area were made in the 1920s, when the close surroundings of the Neusiedler Lake were put under landscape protection. Some 40 years later a provincial law defined the most important areas as fully protected and a few years later the larger area became a man and biosphere reserve. Those formal protection measures did not resolve any of the underlying conflicts and the interests of the landowners remained in contrast to the needs of the protected areas. This was exacerbated by the highly fragmented ownership situation with some 1500 landowners in a relatively small area. An environmental NGO set an early example of successful conservation by negotiations with existing right holders and renting key ecological areas from them. Later governmental institutions followed that example by initiating comprehensive negotiations with existing groups and associations and helped form new groups representing land use rights and concerned communities. This enabled discussions and later schemes compensating for the value of the losses due to restrictions on use. The right to compensation for incurred losses of landowners, hunters and fishermen is now stated in the provincial National Park Act of 1992 (Hubacek and Bauer 1999).

Geobiophysical Scales, Local Institutional Capacity and Institutional Fit

Many 'global' problems are actually local ones with implications on larger regional or

even global scales, such as fragmentation of habitats, deforestation and desertification. Thus many international treaties and multilateral negotiations are ineffective in this case. For example, the Convention on Deforestation will not stop a poor farmer cutting down a tree for firewood (Fraser and Mabee 2002). It is often forgotten that incentive measures and mechanism at the local activities through patchwork addition create larger scale problems. The same applies for the regional and local levels. As Gowdy and Olsen (1992) recognize in their assessment of the deterioration of the environment in the Adirondack Park, NY: “the cumulative impacts of many small unplanned, uncoordinated public and private decisions affecting the natural and cultural resources of the Park are doing as much or more harm to the environment as the impacts of large-scale projects”.

In addition, many regional institutions do not adequately match environmental boundaries—watersheds, pollution distribution depended on wind directions (acidification), migration of species are all compelling examples of environmental realities not adequately captured by institutional set-ups and thus leading to a spatial misfit between institutions and environmental problems. Institutions often exacerbate this with competing agendas creating competing or perverse incentive structures. For example, land drainage programmes of the agricultural sector in Europe in the 1950s through the 1970s were clearly in conflict with the developing environmental regulations and watershed protection acts.

Such a lack of fit causes spatial externalities, benefiting free riders and harming others beyond the reach of the responsible institution. Thus, creating a better fit involves “structuring institutions in ways that maximize compatibility between

institutional attributes and biogeochemical properties through reorganization of political territories and functional cooperation between institutions and jurisdictions (Moss 2004; Young 1999). On the other hand ‘institutional fit’ is an elusive concept that “...can range widely from one situation to another and even from one time period to another with respect to the same institution” (Young 1999).

Institutional fit still seems to apply some notion of top-down legislative land use approaches whereas the previous discussion clearly highlights the importance of public participation, voluntary agreements and inclusions of local knowledge into the decision making processes. One of the most effective means of preserving the environment is for people to take responsibility for the particular place where they live. Therefore, communities need to learn and appreciate the wild patches of their own area (Karasov 1997). To enable people to both learn and value their local environment, appropriate structures need to be implemented to allow information, communication and discussion to flow between locals, politicians, scientists, private initiatives, businesses, and visitors. In other words, the development of institutions and mechanisms to actively involve stakeholders at the local level need to be created. Appropriate institutions should also help to recognize the common good character of many features of land (e.g. biodiversity) and help to find appropriate and viable solutions. This would help insure that people become part of the solution rather than the problem.

Traditional planning and enforcement activities have increasingly been questioned as providing viable long-term solutions for ecosystem destruction and land use conflicts (Brandon and Wells 1992). This new paradigm of stakeholder involvement has found its way into recent policy making and legislation. First experiences show that a

policy style alien to negotiative and participatory governance often poses severe problems of institutional adaptation concluding that “effective implementation is dependent not on the policy type per se but on the degree of congruence—‘fit’—with existing institutional structures and practices” (Knill & Lenschow 2000; Moss 2004). This congruence of cultures and measures is another aspect of the well-known observation that many international environmental regimes require a certain ‘minimum of institutional capacity at lower levels of social organization’ for successful implementation (Young 1999).

Despite some of these problems, there is a growing movement towards greater local participation by empowering a community of stakeholders rather than relying on government officials (Koontz 2003; Kasemir et al 2003). Practical examples include collaborative environmental management and citizen advisory committees or panels. These are usually without binding authority and differ by the extent to which government officials take part. Outcome of such an involvement depend on how the groups organize themselves (e.g. diversity of interest represented, decision rules) and contextual factors such as the level of concern over the issues in the community and pre-existing networks (Koontz 2003:20). One side-effect of such processes might be that in addition to helping solve the problem at hand, collaborative planning leads to an improvement of social capital, interpersonal relationships and networks, on which a community can build upon to solve future problems (Kenney 1999).

In summary, land use decisions often have implications beyond local and regional institutional reach. Decentralized, multi-layered and nested institutions in collaboration with stakeholder networks are one way of balancing the opposing goals of

different stakeholders, and resolving land use conflicts.

Selected References

- Brandon, Katrina Eadie and Michael Wells. (1992) “Planning for People and Parks: Design Dilemmas”, *World Development*, Volume 20, Number 4, pp. 557-70.
- Canada, Government of Canada. (1992) *The State of Canada's Environment*. Ottawa: Government of Canada.
- Coast information Team. (2002) *Coast Information Team*. Government of British Columbia, Canada.
- Delta Farmland and Wildlife Trust. (2000) *Farmland and Wildlife; Grassland Set-Asides* (Fact Sheet #2). DFWT: Delta.
- Delta Farmland and Wildlife Trust. (2001) *Partners in Stewardship* (Promotional Borchure), DFWT: Delta.
- Fraser, E. (2002) “Urban Ecology in Bangkok, Thailand: Community Participation, Urban Agriculture and Forestry”, *Environments*, Volume 30, Number 1, pp. 37-49.
- Fraser, E. (2003) “Land Tenure and Agricultural Management: Soil Conservation on Rented and Owned Fields in Southwest British Columbia”, *Agriculture and Human Values*.
- Fraser, Evan D.G. and Warren Mabee. (2002) “Correspondence: Summit: Vague Answers to Well-Known Problems? Multinational Negotiations Can Work, But Not Where Local People Are Causing the Problem”, *Nature*, Number 418, pp. 817ff.
- Hubacek, Klaus and J. C. J. M. Van Den Bergh. (2002) *The Role of Land in Economic Theory*. International Institute for Applied Systems Analysis: Luxenburg, Austria.
- Hubacek, Klaus and Jose Vazquez. (2002) “Economics of Land Use”, in *Encyclopedia of Life Support Systems*. Paris: UNESCO-EOLSS Publ.

- Hubacek, Klaus and Bauer Wolfgang. (1999) "Austrian Case Study on Economic incentive Measures in the Creation of the National Park Neusiedler See - Seewinkel", in *OECD Case Studies on the Design and Implementation of Incentive Measures for the Conservation and Sustainable Use of Biodiversity*. Working Party on Economic and Environmental Policy Integration. Working Group on Economic Aspects of Biodiversity. Paris: Organisation for Economic Co-Operation and Development.
- Karasov, D. (1997) "Politics At the Scale of Nature", in J.I. Nassauer (Editor), *Placing Nature: Culture and Landscape Ecology*. Washington DC, Coelvo, CA: Island Press.
- Kasemir Bernd, Jill Jäger, Carlo C. Jaeger and Matthew T. Gardner. (2003) (Editors) *Public Participation in Sustainable Science: A Handbook*. Cambridge: Cambridge University Press.
- Kenney, D.S. (1999) Are Community-Based Watershed Groups Really Effective? Confronting the Thorny Issue of Measuring Success. *Chronicle Community*, Volume 3, Number 2, pp. 33-8.
- Knill, C and A Lenschow. 2000 (Editor) *Implementing EU Environmental Policy. New Directions and Old Problems*. Manchester, New York: Manchester University Press.
- Koontz, Thomas M. (2003) "The Farmer, the Planner, and the Local Citizen in the Dell: How Collaborative Groups Plan for Farmland Preservation", *Landscape and Urban Planning*, Volume 66, pp. 19-34.
- Lee, L. (1980) "the Impact of Land Ownership Factors on Soil Conservation", *American Journal of Agricultural Economics*, Volume 62, pp. 1070-76.
- Lumley, S. (1997) "the Environment and Ethics of Discounting: An Empirical Analysis", *Ecological Economics*, Volume 20, pp. 71-82.
- Moss, Timothy. (2004) "The Governance of Land Use in River Basins: Prospects for Overcoming Problems of institutional interplay With the EU Water Framework Directive", *Land Use Policy*, 21, 85-94.
- Olsen, Peg R. and John M. Gowdy. (1992) "Land Use Regulation in the Lake George Basin: An Ecological Economic Perspective", *Ecological Economics*, Volume 6, Number 3, pp. 235-52.
- Ostrom, E. (2001) "Environment and Common Property Institutions", in Neil Smelser (Editor), *International Encyclopedia of the Social and Behavioral Sciences*. Oxford, UK: Elsevier Science, 4-560-66.
- Ostrom, E.; J. Burger; C. Field; R. Norgaard and D. Policansky. (1999) "Revisiting the Commons: Local Lessons, Global Challenges", *Science*, Number 284, pp. 278-82.
- Ostrom, E.; R. Gardner and J Walker. (1994) *Rules, Games and Common-Pool Resources*. Ann Arbor: University of Michigan.
- Praneetvatakul, S.; P. Janekarnkij; C. Potchanasin and K. Prayoonwong. (2001) "Assessing the Sustainability of Agriculture: A Case of Mae Chaem Catchment, Northern Thailand", *Environment International*, Volume 27, pp. 103-09.
- United Nations. (1992) *Agenda 21*. Washington DC: United Nations. www.un.org/esa/sustdev/agenda21.htm
- Young, O. (1999) *Institutional Dimensions of Global Environmental Change*. Bonn: IHDP.

Klaus Hubacek, Evan Fraser
Faculty of Environment, University of Leeds
Leeds. UK
k.hubacek (at) leeds.ac.uk

E.D.G.Fraser@leeds.ac.uk

Shova Thapa

Science and Technology Policy Research Unit

University of Sussex,

Sussex, UK

S.Thapa@sussex.ac.uk

Market for Corporate Control

Ines Perez Soba Aguilar

Introduction

The wave of takeovers that took place in USA in the 1980s revived the analysis of the market for corporate control to such an extent that this has been one of the main topics of discussion in economic and business literature in the second half of the twentieth-century. The market for corporate control is one of the external devices of corporate governance, together with capital, product and factor markets. This is why the analysis of this market is done from the point of view of corporate governance.

Corporate governance can be defined as a set of interconnected institutions and tools that are created with the aim of achieving efficient allocations of present and future resources (Salas 2002). So the characteristics of the market for corporate control and its social function will be related to those of corporate governance.

Models of Corporate Governance

There exist cultural, social, legal, economic, and historical factors that, overtime, strengthen some kind of relationships, not only inside firms but within the society as well (Ménard 1995). Following the differentiation that New Institutional Economics highlights about the sort of relationship established among the parts involved in the activity of a firm (see Putterman, 1986), i.e., a market relationship or a relational one (Goldberg 1980), it is possible to consider two models of corporate governance theoretically: a market model, based largely on market interactions, and a non-market model (or organizational model), which is based on relational exchanges.

Although the sort of model that prevails in real economies is not immutable, since

institutional factors can vary in the long run, from an empirical point of view, taking into account the main institutional variables prevailing in each particular economy, these models have been located in Anglo-Saxon economies (market model) and Continental European and Japanese economies (non-market model). Both models co-exist in market developed economies, so it is difficult to state clearly which one performs better (see Macey 1998 for an evaluation of the relative effectiveness of national corporate governance systems). In fact, several authors (see Mayer 1990) point out that the two models could be in a process of convergence, basing this estimation on the evolution of aggregate data of corporate finance of the last period of the 20th century across developed market economies.

As long as two models of corporate governance can be distinguished theoretically and empirically, two models of market for corporate control can be expected to exist. In this case, the relationship maintained by shareholders in the firm would be the benchmark to differentiate between models. Although there are other very important kinds of interaction with and within the firm that are not materialized in shares, especially in non-market models of corporate governance, when maintaining a relational interaction within the firm, financial and non-financial creditors usually complete their commitment acquiring a position as shareholders. Besides, shares are, up to now, the only financial assets with voting rights that are negotiable in a market. Therefore, when the shareholder that prevails in listed firms is a small, passive and short-term investor, a market model of corporate control is defined. When the main features of the representative shareholder are his large stake and the relational interaction that maintains with the firm, a non-market model of corporate control is thus defined in the economy.

Lately, it has been put the accent on regulation as the key institutional variable to explained differences among ownership structures. Thus, Laporta, López-de-Silanes, Shleifer and Vishny (1998) conclude, basing on the empirical examination of the largest companies of 49 countries, that ownership concentration (that is, large shareholdings) would be a response to poor investor protection.

Definitions

Henry G. Manne (1965) is the first author to study the market for corporate control. Nevertheless, he didn't define that market. Before him, Adolf Berle and Gardiner Means (1932) had analyzed corporate control in a seminal work. Then, Berle in his 1958 paper titled "Control and Corporate Law" summed up the main ideas of that work, and it was this latter work which served as a basis for Manne's well-known paper.

In 1932, Berle and Means distinguished five main forms of control: Control through almost complete ownership, majority control, control through a legal device (e.g. through pyramidal participations), minority control and management control. In 1958, Berle summarized these five forms in just two forms: Absolute or outright control—which includes the three first forms of control stated in 1932—and working control. Under the latter expression Berle groups the ways of providing control *de facto* either on a small shareholder, because of his technique of inducing enough voting stock to elect directors, or on managers, when this substantial minority does not exist and the stock is disperse among many shareholders. This last form of understanding control is the one that has prevailed in the economic literature (particularly, in agency theory) up to the last decade.

Hence, the conventional definition of market for corporate control is that by Jensen

and Ruback (1983): *Market for corporate control is the arena in which alternative management teams compete for the rights to manage corporate resources* (p.42). From the point of view of the models of corporate governance, it could be argued that this definition biases the meaning of this market in two senses: 1. Because it considers only a specific type of control location in the firm: managerial control; 2. Because, besides the theoretical context, it tries to respond to a close reality for them: that of a firm belonging to a market model of corporate governance.

With the intention of defining the market for corporate control in such a way that definition does not vary according to a particular model of corporate governance, it is taken into account the main points discussed by Berle (1958), for whom: a) Control may be defined as the capacity to choose directors and, as a corollary, it carries capacity to influence the board of directors and possibly to dominate it; b) Control is never nonexistent; only its location varies; and c) The function of control is to choose a management. As a result, the following definition is proposed: Market for corporate control is that market in which the capacity of choosing directors directly and managers indirectly is negotiated through transaction of voting shares (or voting rights). Since control rights requires another asset, up to now, to make it tangible and negotiable: shares, we will focus the analysis on shareholders as the main participants on this market.

Characteristics of Markets for Corporate Control

The latter definition of the market for corporate control seeks to be valid for any model of corporate governance, independently of where control is located. The next step is to analyse it through the

characterization of its main components according to each model.

Location of Control

As Berle argues, there are two types of control: the absolute or outright control and the one obtained through an active interest for holding control among a disperse class of passive small shareholders. In the first case, control is in the hands of a large shareholder who, either directly or indirectly, forms the Board of Directors or is responsible for hiring, monitoring and dismissing managers. This situation is common in listed firms of economies in which the non-market model of corporate governance prevails (for example, Germany, Italy, The Netherlands, France or Spain). This shareholder, whether an individual or a legal entity, is, therefore, identifiable, and the absolute (or almost absolute) share of stock he owns increases the probability of negotiating the control of the firm with an acquirer without making a public tender offer to small shareholders. In short, in the non-market (or organizational) model of corporate governance the representative shareholder is a large one who can negotiate the control of the firm without mediation of a public market. A bilateral negotiation between large shareholder and buyer of that control is sufficient.

When large shareholdings don't exist, especially in a situation in which control is in the hands of some managers that enjoy a high degree of autonomy to decide—because the specialization of functions and a low level of monitoring as well—"scenery and protagonists" vary. This is the case prevailing in economies under a market model of corporate governance (for example, United Kingdom or United States). Although there exist minority control or managerial control in the target firm, these are not the indispensable shareholders or agents whom the buyer should negotiate with to get that

control, due to the disperse structure of stockholders. Therefore, the representative shareholder is the small one (where the main source of control is located), and the way to obtain the control is through a public market (e.g. public tender offers).

Demand for Control

The demand for control is made up of those acquirers who want to take control over the firm. In order to get it, they must obtain a "key" percentage of stock firstly. Any individual or legal entity can be part of the demand. Generally, alternative teams of managers have been considered as the "representative buyer" in the market for corporate control, because most of the literature about the market for corporate control is based on the Anglo-Saxon reality, in which the market model of corporate governance prevails. Hence the clarification of Jensen and Ruback (1983:6): "the takeover market is an important component of the managerial labor market", what means that the demand for control is very much related to the supply side of the market of managerial services.

In the non-market model of corporate governance it is possible to find the previous kind of buyer. Nevertheless, in this model, managers are only significant as long as they are at the same time controlling (large or small) shareholders or directors of the target firm. In practice, they do not seem to have a special prominence as bidders in tender offers. The characteristic features of this model are: (a) almost always, it will be indispensable for the acquirer to deal with large shareholder in order to obtain the control of the firm; (b) hence, it is expected that most of the acquirers be of a friendly type, at least for the present group who have the control of the firm, since the concentration of stock and control in the hands of that group would do almost

impossible for acquirer to get the control without their approval; (c) large shareholder of the target firm plays an important role as an acquirer too in his own firm. His goal would be to reinforce his own control to face potential raiders.

Price of Control

The price of the market for corporate control is measured by the premium of control, which allows to allocate the right of control—or the control asset, in terms of Manne (1965)—to the bidder who values it the highest. The price offered for the control is calculated taking into account the value of the firm in ordinary conditions (without an acquisition offering announcement), being the difference of those values the premium of control. Since the market price reflects, from a financial point of view, the present value of shares, the control premium would remunerate the inherent control right that shares contain. Therefore, the value of that premium depends on the value that the buyer and the seller give jointly to that right. From a corporate governance approach, in the economies in which the non-market model prevails, the relevant price in the market for corporate control will be the one given to the large shareholder for this right; and in those economies in which the market model prevails, the relevant price will be the one given to small shareholders.

Free-Rider Problems

According to Manne (1965), a fundamental premise that underlies the market for corporate control is the existence of a high positive correlation between the managerial efficiency in a firm and the market price of its shares. It is supposed that an inefficient listed firm will suffer a fall of its quotation compared with other firms in the same industry. Hence this firm will be more attractive to acquire for those who believe

that, with a more efficient management, it is possible to obtain the latent profit of capital in that firm. If that margin exists, the potential acquirer will be willing to pay for the right to control that firm. Nevertheless, in the well-known article of Grossman and Hart (1980) “Takeovers bids, the free rider problem, and the theory of corporation”, they highlight the difficulty to connect a bad management with the performance of the market for corporate control. This difficulty is due to the following situation: if the shareholders of the target firm, who are supposed to be small shareholders, believe that their decision to sell or not their shares to the acquirer will not have any consequence on the final result of the takeover in a firm with a disperse stockholder structure, they will not sell their shares if its price doesn’t include the future financial benefits achievable by the new team of managers. That is, given that shareholders consider that their individual “vote” lacks importance, only if the premium of control incorporates the present value of those benefits, the shareholders will sell their shares to the acquirer. The problem is that this value discourages potential acquirers from making a takeover bid because it makes negative or nil their benefit.

In this context of certainty and symmetrical information, it could be concluded that, in theory, there would not be any takeover bid. Since Grossman and Hart consider that the externality caused by a good management is what takeover bids try to internalize, they propose, together with other authors who also treat this problem, an array of measures to exclude of that benefit or penalize those who do not incur in the cost to produce it (i.e. the free riders). Some of these measures look the way for achieving to acquirer private benefits to make profitable the takeover, since it would not be possible theoretically for acquirer to earn any financial

benefit. The most important ways are the following:

The first is the dilution of shareholder property rights (Grossman and Hart 1980): In general terms, it consists of setting up the way to create private benefits to acquirers in order to encourage acquisitions at the expense of former shareholders who have not sold. In particular, they propose to reduce the value of the non-tendered shares enough, e.g., arranging to pay statutorily higher salaries to the buyer once he reached the control, or allowing to sell assets of the target firm to the acquirer's firms for a price lower than the market price.

The second is the working involvement of large shareholder (Shleifer and Vishny 1986): The aim would be that large shareholder facilitates the acquirer his entrance in the firm. It could be possible, for example, by splitting with him the gains of the acquisition or, otherwise, by trying not to increase the prices of the offering. They also point out to the possibility that a large shareholder be the acquirer himself.

The third is an extension of the two previous works; it consists of trading the control ("the control block") by means of a private sale between the acquirer and a large shareholder (Burkart, Gromb and Panunzi 1998; Bebchuk 1994), avoiding a public tender offer.

Notwithstanding these measures, it has to be kept in mind that the theoretical analysis of Grossman and Hart is based on the assumptions of certainty and symmetrical information. Since takeover bids are actually carried out, some changes have to be done to accommodate this fact. As Grossman and Hart express, it can be accommodated by relaxing the symmetrical information assumption, introducing different levels of information among the buyer and shareholders, or by incorporating several risk preferences. In general, it is problematic to

value the target firm before an acquisition as a function of the value of the company after it, just as Barnes, Davidson and Wright (1996) argue. In fact, among the parts involved in a takeover, large shareholder, managers and eventually acquirers are the ones who have a better knowledge to estimate the real value of the firm. Small shareholders are, probably, the ones who have less information to calculate an accurate value *ex ante* to be compared to the premium of control offered. In short, the free rider problem—that has so much repercussion in the debate about whether or not legislation must protect small shareholders in takeovers—loses intensity in an environment of uncertainty.

Linkages with other Mechanisms of Corporate Governance

There are three procedures of control (as monitoring)—external, internal and legal—that constrain the movements of a firm: Markets—product and factor markets, financial markets and market for corporate control—are the external ones; monitoring by the board of directors, contracts and financial structure make up the internal devices; and, finally, the legal framework. The performance of the market for corporate control is conditioned by its interrelation with regulation and these other mechanisms of corporate governance.

Given an environment of perfect competition, product and factor markets discipline firms in the long run, whichever the structure of stockholders or control is. But, as the agency theory indicates, there exists a set of mechanisms, prior to the discipline of those markets, which permit to avoid a situation of expulsion from the market. Internal mechanisms, such as contracts or monitoring by the boards of directors, try to prevent manager deviates from the goal of maximization of the shareholders' wealth. In

addition, the managerial labour market (Fama, 1980), the cheapest external solution, would help as a supplier of alternative teams of managers.

If the market for products is perfectly competitive and internal mechanisms of control as well as the managerial labour market fulfil their tasks, the market for corporate control would not be necessary for its disciplinary function. However, natural selection in the product market only discriminates against those firms that do not maximize wealth when there is perfect competition (Winter 1964). Besides, product market discipline works only in the long run. For that very reason, market for corporate control gets its disciplinary meaning (Manne, 1965).

The underlying assumption in Manne's argument is that stock market is efficient, so that stock prices correspond to the present value of the firm. But, when stock market is not efficient and there is not a positive correlation between the managerial efficiency and the market price of the firm, the market for corporate control will not only be unable to guarantee a better reallocation of resources but its function will damage the "natural" selection in product markets, for example, by selecting unprofitable rather than profitable firms for survival (Hughes and Singh, 1989). Chatterjee and Meeks (1996) call attention on to this point, considering that the wave of mergers and acquisitions of the 80s in the United States did not respond to efficient prices. In relation to the economic and social consequences of this takeover wave, there are two versions. Jensen's version (1984) is that the market for corporate control created large benefits for shareholders and for the US economy as a whole by loosening control over vast amounts of resources and by enabling them to move faster to their highest-valued use. On the contrary, for Krugman (1991) all or almost all this gains did not

represent gains to economic efficiency but rather redistribution among stakeholders. Shleifer and Summers (1988) championed this theory, arguing that hostile takeovers facilitate opportunistic behavior at the expense of stakeholders (and their quasi-rents originated in implicit long term contracts), who are expropriated in favor of shareholders.

Apart from Stock Exchange, other financial markets also discipline firms not only through the informative function of prices but through the allocation of financial resources as well. In reference to control (power) of the firm, given that it is negotiated through stock with voting rights, financial creditors usually hold an equity position in order to access to board of directors. Bank shareholdings were a quite common feature in listed firms of non-market model of corporate governance economies in the last century.

Theory and Current Legislation

Since takeover bids have been considered an important tool to improve economic efficiency and, therefore, competitiveness, takeover regulation—along with corporate law, antitrust policy and securities laws—is considered to play an important role in helping to achieve this target. In addition, takeover regulation tries to protect investors, specifically minority shareholders, from being expropriated in the control transaction. A fundamental economic reason to support investor protection is that it is crucial to avoid the reluctance of small investors to participate in the financing of corporations. However, these two goals, i.e. economic efficiency and investor protection, are difficult to be achieved simultaneously because of post-takeover moral hazard by acquirer and free-riding by the target shareholders.

One of the main debates among scholars and regulators alike consists of whether or not small shareholders should be protected by law

when they face a takeover, in order to reach an increase in social efficiency after the transaction of control. Social efficiency is measured by the total value of a firm that a controller can produce with his managerial skills. Thus, a control transaction, i.e. a takeover, will be efficient if the acquirer can produce a higher value in the target firm than before. However, post-takeover moral hazard by acquirer (who can inefficiently extract private benefits from firm value) and free-riding by the target shareholders rise in the way to achieve this goal. Consequently, these two factors are the main topics on which the theoretical discussion about investor protection is posed. In order to analyse optimal strategies of acquirers, the following assumptions are considered: (i) the larger stake the controller has, the more inefficiency he internalizes when extracting private benefits from the firm, and (ii) the target shareholders are free-riders in a takeover. Given these assumptions, two possible optimal strategies can be deduced. Firstly, acquirer will buy as few shares as needed to gain control in order to maximize the value of the control block (the sum of private and security benefits), since free-riders will obtain all the improvement in security benefits that the acquirer will make (Burkart, Gromb and Panunzi, 1998). Secondly, acquirer will buy only the block of control to current controlling party, excluding in the trade all other shareholders, when the shareholder structure is that of a controlling shareholder owning a minority block and the rest of ownership being disperse among small shareholders. This procedure to bargain the control is known as market rule. The market rule, which is largely followed in the United States, maintains that minority shareholders do not have as a whole any right to participate in a control transaction. The other procedure to bargain the control is the equal opportunity rule (EOR), on which some European

countries including Great Britain base their laws, which gives several rights to small shareholders in such sales. Bebchuk (1994) maintains, assuming no correlation between private benefits and the value of the firm, that the market rule is superior to the equal opportunity rule in facilitating efficient transfers of control (because it avoids the free rider problem) but is inferior to it in terms of discouraging inefficient transfers (because of the low level of inefficiency that the acquirer internalizes in these private transactions). According to Burkart, Gromb and Panunzi (1998), assuming a negative correlation between firm value and private benefits, from a social point of view, the market rule is inferior to the equal opportunity rule in any case because it induces more inefficient extraction of private control benefits (or a maximization of the post-takeover moral hazard of acquirer) and the potential negative effect on firm value associated to it, since that rule makes easier for the bidder keeps control with a low level of ownership concentration. So imposing a public tender offer (EOR) improves social welfare. The problem is that, as Berglöf and Burkart (2003) state, when acquirer's private benefits diminish, probability of a takeover diminishes too. For these authors, there is a trade-off between an active market for corporate control and the protection on small shareholders.

Focusing on European and United States regulatory frameworks for the market for corporate control, several differences can be noted (see Forstinger 2002). In the United States, rules are defined by federal legislation (the Williams Act mainly) and State takeover laws, without overlapping their functions. American federal law imposes disclosure requirements on acquisitions above threshold levels or limits the stock holdings of financial institutions, what is considered by several authors as a sign of over-regulation (Roe, 1994), but it doesn't include any takeover

rule. This is the field for State laws whose regulations vary widely from one State to another: For example, Pennsylvania and Maine are known because they are the only states in which there are mandatory bid rules, so the acquirer has to paid all shareholders the same price, and thus small shareholders are supposed to be more protected; in contrast, Delaware is well known because its State law provides incumbents managers excessive protection from hostile takeovers (Bebchuk and Ferrell, 2001).

In Europe, the main goal for European legislator is to achieve the harmonization of takeover regulation in member states, and this target has been probably one of the problematic matters that European Commission has had to face at the turn of the XX century. City Code on Takeovers and Mergers, the self-regulation that United Kingdom introduced in 1968 as a response of its active market for corporate control, may be considered the main source on which other national regulations in Europe and the present European Union (EU) regulation stem from. The essence of City Code is to ensure that in the presence of a takeover operation all shareholders will receive a fair and equal treatment (the equal opportunity rule). The procedure to achieve this goal rests on the information that acquirer has to give to shareholders of target firm and the defensive measures that insiders (managers and board of directors) of target firm are allowed to take when facing a takeover. Besides, in a mandatory offer it is set a minimum limit for the price to be paid to shareholders: it can't be less than any price paid within preceding twelve months.

As Berglöf and Burkart (2003) expound, before the second half of the 1980s, in most Continental European Countries, takeover bids were still so rare that special regulations were long thought to be unnecessary. But later on, as the activity in European market

for corporate control increased, mainly because of the accomplishment of Single Market, most of the countries adopted binding legal rules finally. These regulations are very influenced by City Code, although several peculiarities of each country, especially in the German case, still remain today to such a point that the Directive enacted in 2004 has not included different measures that tried to avoid the power of controlling shareholders to prevent hostile takeovers (e.g., the break-through rule). If it is considered that nowadays regulation is driven by different lobbies, this could be an example of how German lobbies have shown a strong influence within EU. Moreover, this legal result could be explained further by the model of corporate governance that has prevailed for years in each European country. Since that model is generally linked to the origin of the mot representatives companies of the country, the non-market model of corporate governance could be a strong factor to explain how it is possible that to achieve a market-oriented takeover regulation in EU has taken almost thirty years.

Selected References

- Barca, Fabrizio and Marco Becht. (2001) (Editors) *the Control of Corporate Europe*. New York: Oxford University Press.
- Barclay, Michael J. and Clifford G. Holderness. (1989) "Private Benefits From Control of Public Corporation", *Journal of Financial Economics*, 25, 2, 371-395.
- Barnes, Paul; Ian Davidson and Mike Wright. (1996) "the Changing Nature of Corporate Control and Ownership Structure", *Journal of Business Finance & Accounting*, 23, 5/6, 651-671.
- Bebchuk, Lucian A. (1994) "Efficient and Inefficient Sales of Corporate Control", *Quarterly Journal of Economics*, CIX, 439, 957-993.

- Berglöf, Erik (1997) "Reforming Corporate Governance: Redirecting the European Agenda", *Economic Policy*, 24, 93-123.
- Berglöf, Erik and Mike Burkart (2003) "European Takeover Regulation", *Economic Policy*, 173-213.
- Berle, Adolf A. Jr. (1958) "Control" In Corporate Law', *Columbia Law Review*, LVIII, 8, 1212-1225.
- Berle, Adolf A. Jr. and Gardiner C. Means. (1932) *the Modern Corporation and Private Property*. New York: Macmillan.
- Bittlingmayer, George. (2000) "the Market For Corporate Control", in Boudewijn Bouckaert and Gerrit De Geest (Editors), *Encyclopedia of Law and Economics*, Vol. III, the Regulation of Contracts. Northampton, US: Edward Elgar, 725-771.
- Burkart, Mike; Denis Gromb and Fausto Panunzi. (1998a) *Block Premia In Transfers of Corporate Control*. CEPR Discussion Paper, 286.
- Burkart, Mike; Denis Gromb and Fausto Panunzi. (1998b) "Why Higher Takeover Premia Protect Minority Shareholders", *Journal of Political Economy*, 106, 1, 172-204.
- Chatterjee, Robin and Geoff Meeks. (1996) "The Financial Effects of Takeover: Accounting Rates of Return and Accounting Regulation", *Journal of Business Finance and Accounting*, 23, 5-6, 851-868.
- Eggertsson, Thrainn. (1990) *Economic Behaviour and Institutions*. Cambridge University Press, Cambridge.
- Fama, Eugene F. (1980) "Agency Problem and the Theory of the Firm", *Journal of Political Economy*, 88, 288-307.
- Franks, Julian and Colin Mayer. (1990) "Capital Markets and Corporate Control: A Study of France, Germany and the UK", *Economic Policy*, 10, 189-232.
- Goldberg, Victor P. (1980) "Relational Exchange: Economics and Complex Contracts", *American Behavioral Scientist*, 23, 2, 337-352.
- Grossman, Sanford J. and Oliver D. Hart. (1980) "Takeover Bids, the Free Rider Problem and the Theory of the Corporation", *Bell Journal of Economics*, II, 1, 42-64.
- Hughes, Alan and Ajit Singh. (1989) "Takeovers and the Stock Market", In P. Newman; M. Milgate and J. Eatwell (Editors), *New Palgrave: Finance*. London: Macmillan Press, 252-264.
- Jensen, Michael C. (1984) "Takeovers: Folklore and Science", *Harvard Business Review*, 6, 109-121.
- Jensen, Michael C. (1992) "Market For Corporate Control" In P. Newman; M. Milgate and J. Eatwell (Editors), *New Palgrave Dictionary of Money and Finance*. Macmillan Press, London, 657-666.
- Jensen, Michael C. and William H. Meckling. (1976) "Theory of the Firm: Managerial Behaviour, Agency Cost and Ownership Structure", *Journal of Financial Economics*, 3, 4, 305-360.
- Jensen, Michael C. and Richard S. Ruback. (1983) "The Market For Corporate Control: The Scientific Evidence", *Journal of Financial Economics*, 11, 5-50.
- Krugman, Paul R. (1990) *The Age of Diminished Expectations: US Economic Policy in the 1990s*. Cambridge, MA: MIT Press.
- La Porta, Rafael, Florencio López De Silanes, andrei Shleifer and Robert W. Vishny. (1997) "Legal Determinants of External Finance", *Journal of Finance*, 52, 1131-1150.
- Macey, Jonathan R. (1998) "Measuring the Effectiveness of Different Corporate Governance Systems", *Journal of Applied Corporate Governance*, 10, 16-25.

Manne, Henry G. (1965) "Mergers and the Market For Corporate Control", *Journal of Political Economy*, 73, 2, 110-120.

Mayer, Colin. (1990) "Financial Systems, Corporate Finance and Economic Development", in R.G. Hubbard (Editor), *Asymmetric Information, Corporate Finance, and Investment*. Chicago: University of Chicago Press, 307-332.

Ménard, Claude. (1995) "Markets As Institutions Versus Organizations As Markets? Disentangling Some Fundamental Concepts", *Journal of Economic Behavior and Organization*, 28, 2, 161-181.

Putterman, Louis. (1986) *The Economic Nature of the Firm. A Reader*. Cambridge, UK: Cambridge University Press.

Salas, Vicente. (2002) *El Gobierno De La Empresa*. Colección Estudios Económicos, 29, Barcelona: Servicio De Estudios La Caixa.

Shleifer, Andrei and Lawrence H. Summers. (1988) "Breach of Trust in Hostile Takeovers", in A.J. Auerbach (Editors), *Corporate Takeovers: Causes and Consequences*. Chicago: University of Chicago Press, 33-67.

Shleifer, Andrei and Robert W. Vishny. (1986) "Large Shareholders and Corporate Control", *Journal of Political Economy*, 94, 3, 461-488.

Williamson, Oliver. (1985) *Markets and Hierarchies: Analysis and Antitrust Implications*. Free Press, New York.

Winter, Sidney G. Jr. (1964) 'Economic "Natural Selection" and the Theory of the Firm', *Yale Economic Essays*, 4, 1, 225-272.

Encyclopedia of Corporate Governance. www.encycogov.com/A8MarkForCorporateControl.asp.

Corporate Library. www.thecorporatelibrary.com

Inés Pérez-Soba Aguilar
Departamento de Economía Aplicada III
Universidad Complutense de Madrid
Madrid, Spain
iperezso@ccee.ucm.es

Websites

Corporate Governance. www.corpgov.net

European Corporate Governance Institute. www.ecgi.org

Money Laundering

Xiaofen Chen

Introduction

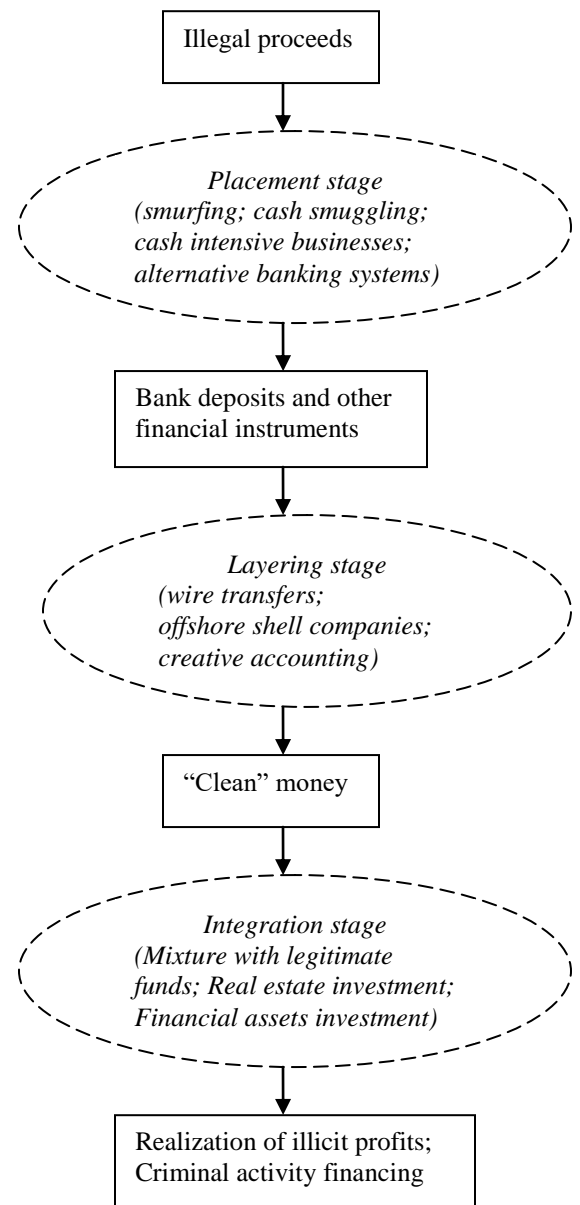
Money laundering is the process of disguising and hiding the origins of illegally obtained funds. Crimes such as drug trafficking, smuggling, robbery, fraud, tax evasion, and corruption generate profits. To avoid being detected, crime offenders “launder” their criminal proceeds (or “dirty money”) so that the funds appear to be generated from legitimate sources and the true origins are non-traceable. Money laundering often involves complex strategies and uses services provided by financial institutions.

A typical process goes through three stages. The first stage is to place illegal proceeds into the financial system. Here, launderers often use a technique called “smurfing”. Large cash deposits are broken into small amounts to avoid being reported. Cash smuggling is also common, especially from the United States to Mexico. After illegal money is smuggled out, it can be deposited in a jurisdiction with lax money laundering controls. The smuggler can also bring it back to the US and acquire proper documents from customs with false invoices. In the second layering stage, funds are moved around through multiple transactions until the origins are erased. At this stage, cross-border wire transfers are often used. The final stage is to integrate funds into legal use. (Gilmore 2004:32; Richards 1999:47-50) Typical techniques used in each stage are shown below in Figure 1.

Money launderers constantly seek alternative channels in response to anti-laundering legislation. When a jurisdiction implements more rigid anti-laundering regulations in the banking sector, launderers usually shift their activities to less regulated financial sectors, such as security trading,

currency exchange, insurance, and investment companies (e.g., venture capital funds and hedge funds). As noted by the FATF (2005), one strategy of using insurance companies is to purchase a single premium life insurance policy and make an early redemption. The FATF (2003) also reports the convenience provided by the international securities market to launder illegal proceeds generated within and outside of the industry. Some of the techniques include cash settlement of security transactions, market manipulation, and the use of a publicly traded company as a front.

Figure 1. Money Laundering Circuit



The trend of using non-financial (i.e., commodity and trade based) businesses to launder money is also growing. As the financial sectors become more rigorously regulated, the traditional use of cash-intensive businesses continues and tends to grow, such as restaurants, liquor stores, car washes, and casinos. In addition, automobile dealers, real estate agents, lawyers, notaries and accountants have also joined the list of professions vulnerable to money laundering activities (FATF 1998, USDS 2005).

The following example described by Richards (1999) and Reason (2001) illustrates how money laundering is completed using a complex network involving financial and non-financial sectors. Columbia is a major supplier of illegal drugs and drug cartels receive massive profits from smuggling drugs to the United States. Typically, to collect them, a Colombian black market peso broker (cambista) is contacted. The cambista instructs his agent in the US to collect the drug money and pays the drug cartel in pesos with a 20 to 40 percent fee. The cambista then uses various strategies such as cash smuggling, front businesses and smurfing to deposit the drug money in a U.S. bank. In the meantime, the cambista sells U.S. dollars at a discount to Columbia importers who want to circumvent exchange controls and avoid paying custom duties when they import goods from the US. The importers pay the cambista in pesos; he then instructs his agent to pay the US companies with drug money through wire transfers. To complete the laundering cycle, goods are smuggled into Colombia, typically via a third country, and secretly sold. According to Richards (1999), 30-40% of U.S. drug proceeds use this type of network.

The Magnitude of Money Laundering

Several approaches can be used to estimate the magnitude of money laundering. Unfortunately, different approaches can yield

widely different results, and no approach is able to produce an accurate measure with confidence. The IMF estimates the total amount of money laundered is about 2 to 5 percent of global GDP (IMF 1998), or between \$0.59 and \$1.48 trillions in 1998. This figure is widely cited by the literature and international organizations, such as the Financial Action Task Force. However, the basis for this estimate is not revealed.

In general, current estimates can be classified into two types: direct estimates using microeconomic approaches and indirect estimates using macroeconomic approaches. Direct estimates are based on suspicious activity reports, sums under investigation, convicted and seized amount, and surveys, etc. An obvious flaw of this approach is that undetected activities cannot be captured. Information generated by surveys is also limited as people are reluctant to reveal sensitive information. Thus, it is impossible to obtain an accurate picture from these estimates. It seems that the estimates are usually biased downward, and sometimes may only reflect a fraction of the actual amount of money laundered.

Two examples of the micro approach are presented by Reuter and Truman (2004) and Walker (1998). In Reuter and Truman (2004), Earnings from 34 crimes in the United States between 1965 and 2000 are estimated to be 6.8-8.1 percent of GDP, of which over half were tax evasion. The top two most profitable crimes were drug trafficking (\$97 billion in 1990) and fraud (\$60 billion in 1990). Walker (1998) estimates the amount of money laundered globally per year is about \$2.85 trillion, based on the average amount laundered per recorded crime and estimated number of crimes. This figure is almost twice the upper bound number provided by the IMF. His work for the Australian Transaction Reports and Analysis Centre suggests between \$1 to \$4.5 billion, or about 0.3-1.3

percent of GDP, is laundered annually in Australia (Walker 1995).

Among the macroeconomic approaches, the currency-demand approach is most popular. First developed by Cagan (1958), and further extended by Tanzi (1980) and Schneider and Enste (2000), the currency-demand approach is used in estimating underground economy activities (including both legal and illegal activities) in a number of countries. As stated in Reuter and Truman (2004:13), the estimate derived from this approach “gives rise to money laundering in its broadest definition”. The approach assumes that a higher tax rate increases incentives for tax evasion and hence the demand for cash. By comparing the difference between the actual cash holding level and the projected level under a low tax scenario, the cash amount associated with the underground economy is derived. The size of the underground economy is then calculated by applying the same income velocity of money for the aboveground economy. Using this approach, Schneider and Enste (2000) estimate the average size of the underground economy amounted 13.5 percent of GDP for the OECD countries during 1990-1993 and grew to 16.9% during 1996-97. The estimates for Australia are 13.0% and 13.9% during these two periods, which are much higher than the estimate by Walker (1995) using a microeconomic approach.

As summarized in Schneider and Enste (2000) and Reuter and Truman (2004), the currency demand approach ignores nonmonetary (e.g., barter), non-cash based (e.g., financial fraud), and non-tax evasion motivated underground activities. It should also be pointed out that some launderers actually pay taxes for the money laundered as part of the laundering process, such as when a front company is used. In addition, a number of the assumptions are flawed. For example, the approach assumes no underground

economic activities exist in the low-tax base year and income velocity of money is the same for both underground and aboveground economies.

Furthermore, the relationship between criminal activities and currency demand is unstable and may change over time. Quirk (1996 and 1997) suggests while an increase in criminal activities may increase currency demand in the 1980s, it may reduce currency demand in the 1990s, probably due to increased application of sophisticated financial market instruments and other non-cash based strategies. More recently, the growing trend of commodity or trade based activities, cash couriers, and informal funds transfer systems (“hawalas”), as a result of more restrictive financial sector regulations (FATF 1998 and USDS 2005), can again reverse the relationship between the two.

There are other less commonly used macroeconomic approaches to measure the magnitude of the underground economy. For example, estimates can be derived from discrepancies between income measure and expenditure measure of national income, between official GDP and the level estimated from electricity consumption, and between official and actual labor force (Schneider and Enste 2000 and Tanzi 1999).

Although no consensus can be reached among various methods, the vast magnitude of money laundering is well recognized. Money laundering has penetrated a wide variety of financial and non-financial industries and increasingly involved professional elites. Mitchell et al. (1998) hypothesize that the corporate culture in the modern capitalist world is a culprit for its emphasis on profits and favor of speculative activities. With competitive pressure, “rule bending” and secret dealings for financial gains are acceptable, while traditional morals are eroded and social responsibilities are lost. In addition, government intervention in

business practices to monitor and investigate money laundering is restricted for ideological reasons. Professions are often trusted to regulate themselves and report suspicious activities, posing a conflict of interest. FATF (2004) reckons that legal and financial professionals have not fully exercised their role as gatekeepers in combating money laundering, due to incentives to maintain client secrecy or lack of awareness. Agarwal and Agarwal (2004) and Masciandaro (1999) have also noted that the culture of secrecy and conflict of interest are obstacles for bankers to cooperate with law enforcement.

Cost of Tolerating Money Laundering

Current analysis of the economic effect of money laundering in the literature is limited to qualitative analysis, probably because it is extremely difficult to quantify. However, there is a growing consensus in the international law enforcement community that the harm caused by money laundering has become intolerable. Considering the underlying effect on the economy, Quirk (1996) suggests the economic cost of money laundering likely exceeds the amount of money being laundered.

A direct result of money laundering is that it facilitates criminal activities. Most criminal activities are for the purpose of gaining profits, and criminal proceeds must be turned into clean money to be usable. Money laundering allows criminals to realize their illegal gains and sustains their incentives to continue and expand their crimes. Widespread criminal, corrupt, and fraudulent activities can paralyze the normal functioning of an economic system and crowd out legal businesses. It is reported that a severe money laundering problem and the associated drug trade, embezzlement, and corruption can significantly hinder economic development and even lead to financial crises, as

experienced by Russia, Mexico and Thailand (Hinterseer 2002:4-7; Fabre 2003:111-134).

One related issue is the debate to decriminalize the drug trade and other vices such as prostitution. For example, Becker et al. (2004) propose to use high taxes in lieu of criminalization of drugs. Indeed, if the drug trade were legalized, drug producers and distributors would no longer need to hide the origins of their profits, and a major source of money laundering would disappear. However, the problem is that the high taxes would provide another incentive for drug suppliers to remain underground. In addition, the price elasticity of demand for drugs may be higher than argued by legalization advocates, as shown in some recent empirical work (for example, Grossman and Chaloupka 1998).

Some observers argue that the inflow of laundered money can create jobs and increase tax revenues, and therefore benefit individual jurisdictions, particularly those lacking investment capital. It also seems that active laundering activities have not caused much harm to offshore financial centers (OFCs. Levi 2002; Alldridge 2003). The counter-argument is that the laundered funds may not be used in a productive way. Money laundering distorts the economic role of financial institutions in allocating resources and creates inefficiency (Quirk 1996). This is because the flow of laundered money is determined by the purpose of hiding the origins and is not directed by interest rates. In addition, the existence of informal funds transfer systems creates parallel exchange and banking markets and can force banks to raise interest rates in order to attract deposits (El Qorchi et al. 2003). Money laundering can also cause disorderly cross-border capital flows, leading to excessive volatility of exchange rates and interest rates and instability of the international financial market (Tanzi 2000).

Furthmore, money laundering erodes the integrity of financial sectors and ruins their reputation. Recently, as the international community becomes increasingly aware of money laundering, OFCs have attracted much attention and are considered the weakest link in combating money laundering. There is no doubt that OFCs can benefit the host economy and the global economy for legitimate uses. However, due to the non-transparent practice of offshore business, inadequate supervision, and bank secrecy laws, OFCs are often used for money laundering purposes. When an OFC gains such a reputation, other jurisdictions may increase scrutiny of its business partners and legitimate customers may avoid doing business with it (see Suss et al 2002:9).

Tax evasion represents a loss of revenue to the government. It undermines governments' ability to provide public goods and hence hinders economic growth. Quirk's empirical work (1996) studies data from 18 industrial countries and indicates that crime is negatively associated with both government spending and growth. However, in terms of the impact of the shadow economy (defined as "legal valued-added creating activities which are not taxed" and violate labor laws, see Schneider and Enste 2000 for a survey), the literature has not reached a consensus. While arguments exist that the shadow economy may benefit countries that lack financial resources and competition and hence may foster growth of the official sector, evidence is also found that the informal sector hurts growth by weakening public infrastructure (Loayza 1997). It should be noted that the channels facilitating tax evasion and criminal activities cannot be disentangled. A complete assessment of the impact of the shadow economy should take this into account. Causality should also be inspected and resources diverted away from the formal economy should be considered to

pinpoint the relationship between underground activities and the official sector.

Another cost of money laundering is related to the fluctuations in currency and money demand caused by shifts of laundering strategies. It can directly create monetary shocks to the economy and increase the difficulty in controlling monetary policy effects, as concluded by Houston (1990).

Lastly, money laundering provides financing channels for terrorist activities. Unlike regular money laundering activities, the purpose of terrorist activities is not to generate profits, and the source of funds can be both legitimate (such as fund raising) and illicit. Nonetheless, it uses the same strategies to hide the ultimate destination of the funds. According to USDS (2005), because terrorist activities require relatively small funds, such financing is easier to accomplish than laundering money in large amounts. Since the 9/11 terrorist attacks, financing terrorist organizations has become a high-profile issue in the US-led international anti-money laundering campaign.

U.S. Anti-Money Laundering Effort

The United States recognized the harm of money laundering relatively early and pioneered anti-laundering legislation. In 1970, it passed the Bank Secrecy Act and obliged banks to retain records and file certain reports for transactions exceeding the specified thresholds. Specifically, banks should file Cash Transaction Reports for all cash transactions above \$10,000, and Currency and Monetary Instruments Reports for cross-border transportation of over \$10,000. The act was to target the underlying crimes (including tax evasion), and did not criminalize money laundering. However, these requirements could not be effectively enforced until the introduction of the amended Bank Secrecy Act in the 1980s. Banks were also asked to report suspicious

transactions since then (Stessens 2000:98; Reuter and Truman 2004:55).

In 1986, money laundering was finally criminalized through the passage of the Money Laundering Control Act. As a result, all financial transactions are considered as money laundering if there is knowledge that the transaction involves proceeds from a “specified unlawful activity” when the transaction is carried out. The list of unlawful activities expanded later and included a broad range of crimes such as drug trafficking, health care fraud, and espionage. Terrorism was included in 1996. However, tax evasion is not in the list, probably due to the resistance from privacy rights advocates and those who maintain that the inflow of foreign funds with a tax evasion purpose is beneficial to the United States. According to Reuter and Truman (2004:54), it is not a problem for the United States to enforce anti-money laundering laws to counter non-tax evasion crimes. Nonetheless, it creates loopholes for foreign residents to evade taxes by investing in the United States and is an obstacle for international anti-money laundering cooperation.

Over the years, a series of new legislation was introduced in the United States, including the 1992 Annunzio-Wylie Money Laundering Act, the 1994 Money Laundering Suppression Act, the 1998 Money Laundering and Financial Crimes Strategy Act, and the 2001 U.S. Patriot Act. The 1992 act included submitting suspicious transactions reports as a mandatory requirement, and demanded banks and other financial institutions to set up internal anti-money laundering programs. The 1998 Act required an annual *National Money Laundering Strategy* to be developed from 1999 to 2003. The 2001 Patriot Act (Title III) closed previous loopholes and strengthened legislation concerning money laundering and terrorist financing. It also provided measures to regulate underground banking systems and

increased powers of relevant government agencies, such as the Internal Revenue Service (IRS) and Financial Crimes Enforcement Network (FinCEN, created in 1990). (Hinterseer 2002:417-438).

Reuter and Truman (2004) reckon that the current U.S. anti-money laundering regime addresses both money laundering prevention and law enforcement. The former aims at deterring money laundering using private resources, while the latter punishes launderers when they are not deterred. Banks, securities firms, and insurance companies are core financial institutions and are subject to most stringent regulations. Less regulated are the non-core financial institutions. They have some reporting duties and may be required to register with FinCEN (such as money service businesses), but do not receive intensive supervision as do core financial institutions. Least regulated are non-financial businesses (e.g., jewelry dealers, real estate agents, travel agencies, and automobile dealers) and professions (e.g., lawyers, notaries, and accountants). In terms of prosecution, launderers are often charged along with their underlying crimes and illegal proceeds are confiscated if convicted.

The responsibilities of supervision and information gathering and processing are shared by FinCEN of the U.S. Treasury Department, the Criminal Intelligence division of the IRS, and the Money Laundering Co-ordination Centre of the Customs Service. Several divisions of the Department of Justice and related state and local agencies are responsible for the enforcement of anti-laundering regimes. A number of other federal agencies are also involved to various degrees (Hinterseer 2002:170).

Global Anti-Money Laundering Effort

Due to the international nature of money laundering, an anti-money laundering effort

has been undertaken by a number of international organizations since the 1980s. An important step taken is the United Nations Convention against Illicit Traffic in Narcotic Drugs and Psychotropic Substances at Vienna in 1988. The Convention took effect in 1990, requiring the signatory countries to criminalize international money laundering and provide international investigation assistance without the constraint of bank secrecy laws. To provide legislation assistance to civil law countries, the United Nations issued the UN Model Law on Money Laundering, Confiscation and International Cooperation in relation to Drugs in 1993. The United Nations Office for Drug Control and Crime Prevention adopted the Global Program against Money Laundering in 1997 to give technical assistance to member countries. (Alldridge 2003).

The most influential international organization in combating money laundering is the Financial Action Task Force (FATF). Established by the G-7 Summit in Paris in 1989, the FATF carries the mission of combating money laundering with 16 original members. It later expanded to 33 members (including 2 regional institutions), and facilitated the establishment of regional bodies in the Asia-Pacific, Caribbean, Europe, Eastern and Southern Africa, and South America to ensure the implementation of its recommendations.

Over the years, the FATF has made significant contributions in setting up the global anti-money laundering framework and fostering international cooperation. One of its major contributions is the issuance of the Forty Recommendations for Combating Money Laundering in 1990. Although not binding, the Recommendations were widely adopted by international organizations and served as the principles for individual jurisdictions to develop their own legal frameworks. These recommendations were

revised in 1996 and again in 2003 to keep pace with the development of money laundering trends. After the 2001 terrorist attacks in the US, the FATF included combating terrorist financing as one of its purposes and issued the Eight Special Recommendations on Terrorist Financing.

In 2000, the FATF launched the process of identifying the non-cooperative countries and territories and published the first list of 15 jurisdictions that had not complied with the recommendations based on 25 assessment criteria. It was recommended that financial transactions with the uncooperative jurisdictions should be subject to increased scrutiny. This “name and shame” process has been successful in urging countries in the list to speed up the process of implementing and improving their anti-money laundering measures. There are now no jurisdictions on the list.

Other international organizations, such as the Council of Europe, the Basle Committee, the World Bank, and the International Monetary Fund (IMF) are also involved in promoting anti-money laundering enforcement and providing relevant assistance. For instance, the IMF launched a program of assessing offshore financial centers in 2000 and later incorporated the program as a standard part of its operation, even though many of the offshore centers are not IMF members. The program developed a comprehensive methodology to assess the adequacy of bank supervision and measures of combating money laundering and terrorist financing in consistency with FATF’s recommendations. It finds that smaller and poorer jurisdictions have lower compliance levels and their anti-money laundering regimes need to be strengthened. Of the 41 jurisdictions it assessed, about 40% had inadequate anti-money laundering rules for branches and subsidiaries located abroad and 50% needed to improve measures of

combating terrorist financing (IMF 2004). During the assessment process, many jurisdictions, particularly poorer ones, received technical assistance and made significant progress by taking measures such as setting up financial intelligence units (FIUs) to gather and process information on financial crimes (Darbar et al 2003).

The European Union passed the Money Laundering Directive in 1991 and required that member states should ensure money laundering is prohibited. Credit and financial institutions are obligated to collect customer identification information for transactions above ECU15,000 and report suspicious transactions (Directive 91/308/EEC). However, the Directive does not have specific definition to cover crimes other than drug trafficking. As a remedy, the 2001 Amending Directive broadened the coverage of criminal offences and included those defined in the Vienna Convention and other “serious offenses” (Directive 2001/97/EC). It also mandated professions such as auditors, tax advisors, lawyers, accountants, notaries, auctioneers, and real estate agents, etc., to be subject to the 1991 Directive. The Third Directive was proposed in 2004 and approved in 2005 to address the revised recommendations issued by the FATF in 2003. The Directive will apply to all providers of goods for cash payments over €15,000. The deadline for completing the implementation of the Directive has long past (December 2007).

Many EU countries set up a FIU with similar functions as FinCEN in the United States. However, unlike its EU counterparts, FinCEN has unique functions in regulating and supervising banks. Across European Union, the nature of FIUs also differs. Some are part of the police force (e.g., Switzerland) or judicial authorities (e.g., Luxembourg and Denmark). Some are supervised by the central bank or the treasury (e.g., France and

Italy), but are relatively independent. A few are completely independent (Belgium and the Netherlands). Depending on its nature, the FIUs are subject to restrictions on disclosing information to other authorities to various degrees. (Stessens 2000:183-199).

Reuter and Truman (2004:90-91) observe that the approaches in criminalizing money laundering differ in the United States (in connection with specific crimes) and most other countries (in connection with all “serious offenses”), such as the European Union. They comment that each has its pros and cons. In an effort to bridge the two approaches, the 2003 FATF Recommendations gave 20 broad categories of predicate offenses. However, the list did not include tax evasion, even though some countries (e.g., France) incorporated it in their national legislation. As criticized by the authors, the lack of uniform legislation across jurisdictions may hinder international cooperation. Similarly, Mitsilegas (2003) indicates that there is evidence of a “double standard” and inconsistency in assessing FATF member and non-member jurisdictions, causing confusion and reluctance to cooperate between nations.

Improving Global Anti-Money Laundering Regimes

With the progress of the global anti-laundering regime accelerating, a large number of articles and books have been published on this issue recently. However, there is a clear need for further research to provide more background information for policy making. For example, the literature has not reached a consensus on assessing the magnitude of money laundering and quantifying its micro and macro impact. Most of the existing research studies the underground or shadow economy. There is very little literature directly targeting money laundering on this issue. In order to achieve

the purpose, better data as well as consistent international evaluation methodologies are required (Quirk 1997).

More information is also needed to assess the costs and effectiveness of the current anti-money laundering regulations. Reuter and Truman (2004) suggest that the costs include three types. First, it requires public resources to set up and operate government agencies to process information, supervise compliance, and investigate and prosecute violations. Second, private institutions also incur costs to comply with regulations. As reported by Bosworth-Davies (1997), a survey of U.K. money laundering reporting officers suggests that training staff and obtaining client identification incur highest compliance cost. A third type of costs is borne by the general public. Based on existing relevant work, Reuter and Truman (2004) estimate the U.S. anti-laundering regime roughly costs \$7 billion per year. As they point out, more accurate estimates are needed for the United States and other jurisdictions. Masciandaro (1999) also concludes that regulation costs pose constraints on the strictness of anti-laundering controls and determine the degree of money laundering tolerance. This is an issue that must be considered in all jurisdictions. Poorer jurisdictions may lack the resources to enforce similar laws as in richer jurisdictions, creating difficulties in international cooperation.

On the other hand, the effectiveness of anti-money laundering laws also needs to be evaluated. Some provisions may need to be reconsidered. For example, there have been concerns that the current reporting duties of U.S. financial institutions produce too much information, delaying the processing of valuable information and reducing the effectiveness of the system. Thus, it was recommended that the threshold for banks to file Currency Transaction Reports be raised (Reuter and Truman 2004:55). On the other

hand, there is also evidence that the threshold is too high to effectively combat terrorist financing (Hinterseer 2002:414).

Obstacles also exist in enforcing anti-money laundering laws. As mentioned above, there is a clear need to harmonize anti-money laundering legislation. Hinterseer (2002) points out that prosecution also requires international cooperation. Even though there have been international conventions addressing the jurisdiction issue, launderers may still escape punishment if the involved jurisdictions are reluctant to cooperate. National blocking laws and bank secrecy laws prevent information from being disclosed to authorities and other jurisdictions if there is no anti-laundering legislation to override these laws. There also exists an incentive issue for private institutions to cooperate. For example, even when banks are required to cooperate with authorities, they may lack the incentives to voluntarily disclose information regarding their clients, either as a long established tradition to protect the interests of their clients, or due to concerns that their reputations may be damaged.

Selected References:

- Agarwal, J.D. and Aman Agarwal. (2004) "International Money Laundering in the Banking Sector", *Finance India*, 18, 2, 767-778.
- Alldrige, Peter. (2003), *Money Laundering Law: Forfeiture, Confiscation, Civil Recovery, Criminal Laundering and Taxation of the Proceeds of Crime*. Oxford: Hart Publishing.
- Becker, Gary S.; Kevin M. Murphy and Michael Grossman. (2004) *The Economic Theory of Illegal Goods: The Case of Drugs*. New York: National Bureau of Economic Research Working Paper 10976.
- Bosworth-Davies, Rowan. (1997) *The Impact of International Money Laundering*

- Legislation*. London: FT Financial Publishing.
- Cagan, Phillip. (1958), "The Demand for Currency Relative to the Total Money Supply", *Journal of Political Economy*, 66, 4, 303-328.
- Darbar, Salim M.; R. Barry Johnston and Mary G. Zephirin (2003) "Assessing Offshore: Filling a Gap in Global Surveillance", *Finance and Development*, 40, 3 (September), 32-35.
- El Qorchi, Mohammed; Samuel Maimbo and John F. Wilson. (2003) *Informal Funds Transfer Systems: An Analysis of the Informal Hawala System*. Washington DC: International Monetary Fund Occasional Paper 222.
- Fabre, Guilhem. (2003) *Criminal Prosperity: Drug Trafficking, Money Laundering and Financial Crises After the Cold War*. London: RoutledgeCurzon.
- FATF. (Financial Action Task Force on Money Laundering) (2005) *Report on Money Laundering Typologies, 2004-2005*. Paris: FATF.
- FATF. (Financial Action Task Force on Money Laundering) (2004) *Report on Money Laundering Typologies, 2003-2004*. Paris: FATF.
- FATF. (Financial Action Task Force on Money Laundering) (2003) *Report on Money Laundering and Terrorist Financing Typologies, 2002-2003*. Paris: FATF.
- FATF. (Financial Action Task Force on Money Laundering) (1998) *1997-1998 Report on Money Laundering Typologies*. Paris: FATF.
- Gilmore, William C. (2004) *Dirty Money: the Evolution of International Measures to Counter Money Laundering and the Financing of Terrorism*. Council of Europe, Strasbourg.
- Grossman, Michael and Frank J. Chaloupka. (1998) "The Demand for Cocaine by Young Adults: A Rational Addiction Approach". *Journal of Health Economics*, 17, 4 (August), 427-74.
- Hinterseer, Kris. (2002) *Criminal Finance: the Political Economy of Money Laundering in a Comparative Legal Context*. Hague: Kluwer Law International.
- Houston, Joel F. (1990) "The Policy Implications of the Underground Economy", *Journal of Economics and Business*, 42, 1 (February), 27-37.
- IMF (International Monetary Fund). (2004) *Official Financial Centers: The Assessment Program—An Update*. March 12. Washington, DC: IMF.
- IMF (International Monetary Fund). (1998) "Money Laundering: the Importance of International Countermeasures", by Michel Camdessus, IMF Managing Director. Washington, DC.: IMF. www.imf.org/external/np/speeches/1998/021098.htm
- Levi, Michael. (2002) "Money Laundering and Its Regulation". *Annals of the American Academy*, 582, July, 181-194.
- Loayza, Norman V. (1997) *The Economics of the Informal Sector: A Simple Model and Some Empirical Evidence from Latin America*. Working Paper 1727. Washington DC: World Bank.
- Masciandaro, Donato. (1999) "Money Laundering: the Economics of Regulation", *European Journal of Law and Economics*, 7, 225-240.
- Mitchell, Austin; Prem Sikka and High Willmott. (1998) *The Accountants' Laundromat*. Essex, UK: Association for Accountancy & Business Affairs. visar.csustan.edu/aaba/laundry.htm
- Mitsilegas, Valsamis. (2003), "Countering the Chameleon Threat of Dirty Money: 'Hard' and 'Soft' Law in the Emergence of a Global Regime Against Money Laundering and Terrorist Finance", in

- Adam Edwards and Peter Gill (Editors), *Transnational Organised Crime: Perspectives on Global Security*. London and New York: Routledge, 185-211.
- Quirk, Peter J. (1997) "Money Laundering: Muddying the Macroeconomy", *Finance & Development*, 34, 1 (March), 7-9.
- Quirk, Peter J. (1996) *Macroeconomic Implications of Money Laundering*. International Monetary Fund Working Paper 96/66. Washington DC: IMF.
- Reason, Tim. (2001) "The Corporate Connection: How Drug Money is Finding its Way to the Bottom Line", *CFO Magazine*. www.cfo.com/magazine/index.cfm/3046501
- Reuter, Peter; and Edwin Truman. (2004) *Chasing Dirty Money*. Washington DC: Institute for International Economics.
- Richards, James R. (1999) *Transnational Criminal Organizations, Cybercrime, and Money Laundering*. Boca Raton, Florida: CRC Press.
- Schneider, Friedrich and Dominik H. Enste. (2000) "Shadow Economies: Size, Causes, and Consequences", *Journal of Economic Literature*, 38, 1 (March), 77-114.
- Stessens, Guy. (2000) *Money Laundering: A New International Law Enforcement Model*. Cambridge, UK: Cambridge University Press.
- Suss, Esther C.; Oral H. Williams and Chandima Mendis. (2002) *Caribbean Offshore Financial Centers: Past, Present, and Possibilities for the Futures*. International Monetary Fund Working Paper 02/88. Washington DC: IMF.
- Tanzi, Vito. (2000) "Money Laundering and the International Financial System", in *Policies, Institutions and the Dark Side of Economics*. Cheltenham, UK: Edward Elgar, 186-200.
- Tanzi, Vito. (1999) "Uses and Abuses of Estimates of the Underground Economy", *Economic Journal*, 109, 456, F338-F347.
- Tanzi, Vito. (1980) "The Underground Economy in the United States: Estimates and Implications", *Banca Nazionale del Lavoro Quarterly Review*, 135, 428-453.
- USDS. (U.S. Department of State) (2005) *International Narcotics Control Strategy Report, II: Money Laundering and Financial Crimes*. Washington DC: Bureau for International Narcotics and Law Enforcement Affairs.
- Walker, John. (1998) *Modeling Global Money Laundering Flows*. John Walker Crime Trends Analysis. www.johnwalkercrimetrendsanalysis.com.au
- Walker, John. (1995) *Estimates of the Extent of Money Laundering In and Throughout Australia*. Sydney: Australian Transaction Reports and Analysis Centre.

Websites

- FATF. (Financial Action Task Force) www.fatf-gafi.org
- International Monetary Fund. *The IMF and the Fight Against Money Laundering and the Financing of Terrorism*. www.imf.org/external/np/exr/facts/aml.htm
- United States Department of the Treasury. Financial Crimes Enforcement Network. www.fincen.gov

Xiaofen Chen

Department of Economics
Truman State University
Kirksville, Missouri, USA
xiaofen@truman.edu

Neo-Malthusianism, Population and Environment

*Evan Fraser, Klaus Hubacek
and Katarina Korytarova*

Introduction

Over two hundred years ago, the Reverend Thomas Malthus made a seminal contribution to explaining population growth and hunger in his *Essay on Population* (1798). Observing the fast-growing Irish population, he argued that a limited amount of agricultural land and high population growth would inevitably lead to hunger, famine, disease and death: "...population, when unchecked, increases in a geometrical ratio... [while] subsistence increases only in an arithmetical ratio..." (Malthus 1976:ch 1). Coming at a time when famines ravaged much of the colonial world (in addition to the Great Irish Potato Famine that claimed 1 million lives and forced another 1 million into exile, between 1876 and 1902 approximately 30 to 60 million people died of hunger in India, China and Brazil (Davis 2001)) the "Malthusian argument" provided a justification for political inaction. Lord Lytton, Viceroy of India during a catastrophic late 19th century famine, remarked that the calamity in India was caused by the Indian population's "...tendency to increase more rapidly than the food it raises from the soil..." (Quoted in Davis 2001:32).

Malthus' theory provided a 'natural law' for inequality and the misery of the masses, and was immediately very influential. For example, Charles Darwin (1809-1882), author of *Origin of the Species* (1859), was directly influenced by Malthus and wrote in his autobiography (published in 1876):

"In October 1838, that is, fifteen months after I had begun my systematic inquiry, I happened to read for amusement Malthus on Population, and being well prepared to

appreciate the struggle for existence which everywhere goes on from long-continued observation of the habits of animals and plants, it at once struck me that under these circumstances favourable variations would tend to be preserved, and unfavourable ones to be destroyed. The results of this would be the formation of a new species. Here, then I had at last got a theory by which to work."

Malthus' ideas were enthusiastically received by Victorian policy makers reflecting "...the overall societal situation and mind of the industrializing Victorian England" (Seidl and Tisdell 1999:397). Policy makers who accepted the Malthusian argument as sound, viewed famine as an entirely natural – though regrettable – way to bring humanity back into balance with the environment. This political opinion was driven by the invisible hand of market economics mixed with the Social Darwinian assumption that the British Empire ruled the world due to the superiority of its value system (Woodham-Smith, 1962) for an investigation of the Irish Famine in light of British Politics). Any attempt to mitigate the suffering caused by famine risked being attacked in London as making the problem worse: "Every benevolent attempt made to mitigate the effects of famine...serve but to enhance the evils resulting from overpopulation" (Sir Evelyn Baring, then Finance Minister, quoted in Davis 2001:32).

The logic is seductively simple: poor primitive societies lacked the moral code required for self-restraint and bred until they overwhelmed the environment's capacity to sustain bloated populations. Famine was nature's inevitable way of curbing excess populations. As such, policies designed to stave off the effects of famine (such as establishing public works, selling low cost food into local markets, and providing direct food aid) served to further inflate populations

beyond a sustainable level and delay the inevitable day when populations would crash.

Modern assessments tell a very different tale. Each of these “Victorian” famines was precipitated by an environmental trigger (the outbreak of the potato blight in the case of Ireland and El Nino induced drought in the case of India, China and Brazil). The effects of the famine were then exacerbated by a host of socio-economic factors that prevented starving people from accessing the food that was available. Population growth seems to have little to do with the problem, and in Ireland it was poor and isolated communities that suffered not those with high population growth rates (Fraser 2003). While it is an exaggeration to suggest that the Colonial powers did nothing to ease the pain (O’Grada 1989), relief efforts were hampered by those who had their worldview confirmed by Thomas Malthus’ logic.

Furthermore, Malthus ignored the extremely important role that innovation, technology and ingenuity play in increasing food production. Put simply, Malthus was convinced that the demand for food would inevitably outstrip the supply of food. In the last 200 years, the reverse has been true. In almost every region around the world, food production has grown faster – sometimes much faster – than population growth. This has resulted in healthier and longer-lived human populations in most regions.

The influence of Malthus’ ideas have waxed and waned over the decades. More recently the original flavor of the Malthusian apocalypse reemerged with Milton Keynes (1920) who said “...the ‘Malthusian devil,’ chained for more than half a century, was unleashed again.” Keynes was referring to the period during and after World War I, when supplies of food and fibre could not keep up with demand (quoted in Ely and Wehrwein, 1948:10).

Garrett Hardin, author of *The Tragedy of the Commons*, uses Malthusian conclusions, when he suggests that any attempt to help the poor will result in a situation in which, the less provident and less able, will multiply at the expense of the abler and the more provident, bringing eventual ruin upon all who share in the commons (Hardin, 1974). Malthusian logic was also applied to early computer models that attempted to anticipate the future of the planet in light of rising population. The best known example of these models is *The Limits to Growth* (Meadows and Club of Rome 1972).

Population Growth and Carrying Capacity

Malthus, often considered “... a ‘philosopher’ who first saw the importance of the limiting factor of environment on human material process” (Bowen, quoted in Seidl and Tisdell, 1999:397), also had enormous influence on the development of one of the cornerstone concepts in ecology: *carrying capacity* (ibid p.396) Carrying capacity is the theoretical maximum population that an area can sustain under given technological capacities and natural constraints. The notion of carrying capacity underpins much of what we now call sustainable development, which has been defined as a necessary and sufficient condition for a population to be at or below carrying capacity (Daily and Ehrlich 1992).

Despite the intuitive clarity of the concept, a debate has developed on how to measure carrying capacity. Ecologists sometimes define carrying capacity based on Justus Freiherr von Liebig’s “*law of the minimum*.” The law of the minimum asserts that under reasonably constant and stable conditions, the population size of any species is constrained by whatever resource is in the shortest supply. Originally, this law was devised to account for plant growth in environments where different nutrients were limited (Liebig 1859). It has been subsequently applied to human

populations (Cohen 1995a). For example, Brian J. Skinner, a geologist at Yale University, claims that “More than any other factor, the availability of water determines the ultimate population capacity of a geographic province” (Cohen 1995a). The availability of food, energy, land, soil, space, diseases, waste disposal, and non-fuel minerals, have all been proposed as potential natural constraints that could impose limits on human populations (Cohen 1995a:329).

Over time, projections of carrying capacities have not only been based on how single resources constrain populations but also how numerous different and interacting resources can affect population. The most recent and popular version of this sort of thinking is the *ecological footprint* that tries to relate the full impact of consumer behaviour to the amount of land taken up by these actions. This concept was established by Wackernagel and Rees (Rees and Wackernagel, 1995) who reasoned that almost everything we do can be related to land use: land used for crops, land used for transportation, land used for garbage disposal, and even land that should be used to plant trees to sequester carbon emitted by burning fossil fuels.

Dauvergne (1997) extends this argument further by introducing the notion of the “shadow ecology” to represent the impact that many of our activities have on distant places. This concept allows us to identify the remote sources of environmental pollution, even if they come from companies who are based in corporate headquarters in different countries. Findings based on the ecological footprint or shadow ecology models suggest that rich, powerful or developed nations maintain their privileged place in the world (and relatively pristine environments) partly by transferring many of the negative consequences of industrialization onto poorer regions (e.g.

MacNeill et al 1991, Giljum and Hubacek 2001).

Another way of assessing the impact of population on the environment was initiated by the debate between Paul Ehrlich, Jon B. Holdren and Barry Commoner in the early 1970s. At the crux of this famous exchange were questions about whether population growth, levels of consumption or technology were the most important factors driving the degradation of the global environment (Ehrlich 1970; Commoner 1971). These scholars developed a simple analytical tool to assess the contribution of population growth, the amount of resources people consume and the types of technology they use, that has been formalized in the following equation:

$$I = P \times A \times T$$

In this equation I = the environmental impact of a community, P = the size of the population, A = their affluence, and T = the technology they use. This equation has been frequently applied and modified (e.g. Fischer-Kowalski and Aman 2001; York 2003), but the debate has never been adequately resolved as to whether or not population growth is the most imminent problem as originally claimed.

Although simple, the IPAT approach is more sophisticated than Malthus', as it widens the discussion from a narrow focus on population to questions of lifestyles and consumption. This complexity is aptly expressed by Joel Cohen:

“How many people Earth can support depends in part on how many will wear cotton and how many polyester; on how many will eat meat and how many bean sprouts; on how many will want parks and how many will want parking lots. These choices will change in time and so will the number of people Earth can support.” (Cohen 1995b).

Similarly, Rao (2000) looks at the amount of energy use around the world, and suggests that in terms of impact, one American born in

1975 uses the same resources as the birth of 50 Indians born at the same time. Rao uses this assessment to conclude that policies aimed at reducing birth rates in the developing world only divert attention from the much more significant problem: resources are exploited in the third world by the first world nations and that there is a net transfer of resources from the developing world to the industrialized world (Rao 2000). Despite these observations policies have continued to focus on the population question.

Neo-Malthusiasm and Population Policies

As early as 1891, the revived Malthus ideas of overpopulation served as an explanation for poverty in developing countries. Neo-Malthusiasm, movement formed in the late 19th century influenced by eugenics and birth control movement, drove policy focused on control of the population size, which was seen as a barrier to development. In India this movement spurred organizing the First Family Hygiene Conference (1938) and founding International Planned Parenthood Federation (IPPF Bombay) in 1952, which soon became a major force in the population control movement across the globe. IPPF got increasing funding from Hugh Moore Fund and Rockefeller Foundation, major corporations as well as by US government. President Jackson justified this funding by stating: "Let us act on the fact that less than five dollars invested in population control is worth a hundred dollars invested in economic growth" (Jackson 1965, cited by Rao 2000). Similarly, the World Bank stated: "rapid growth of population has become a major obstacle to social and economic development in many of our member states. Family planning programs are less costly than conventional development programs (Mass 1979, cited by Rao 2000).

In 1952, India initiated one of the first family planning programs in the world

introducing massive family control measures that included vasectomies and female sterilization. Subsequently, it was recognized this had failed (Joshi 1974, cited by Rao 2000). One of the chief reasons cited was that the government was not able to recognize that motivation to practice family planning is dependent on the socio-economic situation of parents, which in turn alters the determinants of family size. While neo-Malthusiasm holds that the people are poor because they have large families, it seems that the poor require larger families because they are poor (Mamdani 1973, cited by Rao 2000), and that it is the children who help them provide income and food.

Approaches to Food Security

Another area where Malthusian thinking has had an enormous impact is in the field of food security, which is defined by the World Bank as the state when all people have access at all times to a stable and nutritionally adequate food supply. When Malthusian and Neo-Malthusian theories are applied to food policy, they are similar to a great many development theories popular in the twentieth century in that they make gross generalizations about the way the world works, assuming that the same set of conditions leads to the same outcomes, regardless of local specifics. For example, it was quite popular in the mid twentieth century to assume that all countries would progress along the same trajectory from a "traditional" economy to a modern one (e.g. Rostow 1971). However, closer inspection reveals that the world is far more complicated. The same types of agricultural technology that helped North America to increase yields (mechanization) simply exacerbated rural unemployment in South East Asia (Bray 1986; Scott 1985). As a result, there has been a considerable backlash against these sorts of global development

theories and contemporary scholarship tends to emphasize the need to understand local factors: The defence of the local as a prerequisite to engaging with the global; the critique of the group's own situation, values and practices as a way of clarifying and strengthening identity; the opposition to modernising development; and the formulation of visions and concrete proposals in the context of existing constraints, these seem to be the principal elements for the collective construction of alternatives that these [Third World] groups seem to be pursuing (Escobar 1995:226).

In reaction to the failures of these large-scale approaches, a different way of looking at food security has emerged over the last twenty years. Rather than being driven by global theories about the relationship between population and hunger, policy has emerged from the theory that hunger is a local problem that can only be addressed at the individual level. The methods to collect the information that is supposed to help policy makers understand hunger fall into two categories. The first is generally termed "non-welfare" or "non-hedonic." Food security policies based on non-hedonic methods are directed by data that attempt to categorize people as nutritionally poor if they fail to achieve certain health standards (say by seeing if children are below a certain weight by a given age). This is in contrast with policies that are directed by "hedonic" or "welfare" methods that ask the people themselves to describe their own welfare.

The non-hedonic position makes no attempt to identify the causes of hunger. Rather, according to this school, the key is to set basic health standards and then assess peoples' access to food on this basis. In this way, if a person is deficient in a certain micronutrient then health officials will be able to provide that micronutrient.

Unfortunately, this process is fraught with methodological problems. Every way of assessing nutritional deficiencies is undermined by some practical consideration that makes this whole approach problematic (Foster 1992). Based on this approach, policy-makers are equipped with nothing more than raw data that is of questionable usefulness. For example, nutritional problems, health deficiencies, and stunted development are all symptoms of hunger and malnutrition. Merely identifying the symptoms will provide no guidance on how to solve this problem. As a result, policies that emerge from non-hedonic methods will fail to take into account the causes of hunger.

In contrast, hedonic methods make comparisons of welfare and public policy based on the preferences of individuals. This approach is based on the idea that people will have a "preferential ordering of goods" that represents a "utility function" (Ravallion 1994:4). In other words, people are able to recognize what is useful to them, and will choose those things. In this way, you could present a community with a number of different policies and let them decide the most appropriate one for their situation. It is simply a matter of asking people what they want and whether they have the means to obtain it.

But there are also a number of problems associated with hedonic approaches. First of all, this method for assessing poverty is largely theoretical and relatively untested (Paim 1995). Secondly, people may not always be the best judge of their welfare. Thirdly, the pursuit of individual welfare may not enhance the welfare of the larger community. For example, in a situation where resources are scarce it is often in the best interests of the individual family to have many children so that they can capture a larger share of the economy. If, however, all families follow this strategy, there may

ultimately be fewer resources and may thus decrease everyone's well-being. There is an additional problem of "asymmetrical welfare distribution" within the family. If a family does not share resources equally, then a policy that centres on the family may fail to address the needs of individual family members.

The concern that not all families share resources is backed up by a series of empirical studies. For example, one study on single-parent households in Africa showed that an increase in the income of female-headed households increased the household's food, health, and education budgets by 3-6 times more than if the same income was given to male headed households (Haddad et al 1997). According to this study, if the single parent is male, there is less distribution of additional income than when the household head is female. Another study from Africa echoes this conclusion: Keopman (1997) illustrates that in most situations in rural Africa, household incomes are not generally pooled. Rather, women tend to be responsible for food, while men are generally responsible for housing for the family (Haddad et. al 1997:130). Traditional methods of poverty assessment, which rely on the household as the basic economic unit in a society may not adequately explain the complexities of gender relations in many societies.

There is a third way for assessing food security that avoids the methodological problems just outlined. This approach focuses on the economic ability that an individual or single family has to deal with their own problems. The most famous proponent of this approach is the Nobel laureate, economist Amartya Sen, who argues that the study of food security should focus on people's capability to obtain food. Sen defines capability as the ability to undertake specific objectives that are useful to the family (Sen 1981,1987). Ravallion (1994) examines

people's economic capability to obtain food when he defines poverty as occurring "... in a given society when a person(s) does not attain a level of economic well-being to constitute a reasonable minimum by the standards of that society." (p. 3). Alternatively, it is possible to determine what percent of a family's budget is spent on food. The economist's job is to then determine what can safely be spent on food while leaving sufficient money left over to purchase the rest of the family's needs such as shelter and education. Based on the cost of living in a society, if a family, individual or group's food expenditures within a given budget grew too large, they would be considered food insecure.

The benefit of looking at food security from the perspective of economic capability is that policy makers will not pre-suppose how people should be living or the types of food people should be eating. Rather, this approach just measures how much freedom an individual or family has.

The disadvantage of this approach is that (a) some people--especially women--do not work for a wage, so it is very important to study more than just economic activity; and (b) when faced with hunger some people will not only use cash to find food. People may switch to inferior foods, cease waged labour and return to subsistence production. A simple poverty-line approach may not pick up these very important aspects of food security. As a result, Sen also coined the term *food entitlement* to describe the many different ways in which a group obtains food: the failure to obtain food, therefore, becomes an "entitlement failure" and can occur anywhere between the producers and the consumer of food. To fully understand food entitlements, it is necessary to measure economic assets (such as money in a bank account) but also human, social, and natural capital too: In each social structure, a person can establish command over some alternative commodity

bundles (any one bundle of which he or she can choose to consume) ... The set of alternative bundles of commodities over which a person can establish command will be referred to as this person's entitlement (Dreze and Sen 1989, p. 5).

For example, the way a person obtains food (or achieves their "entitlement") can come from either direct sources (e.g. a farming family that grows its own food), indirect sources (e.g. a labouring family that exchanges money for food and obtains a regular income) or transfers (e.g. charity and food aid). Acute malnutrition and famine occur when a person's or a community's "entitlement" is disrupted. This can be an indirect or demand-side failure, which occurs when people lose their purchasing power through unemployment, falling wages, rising food prices, or inflation and do not have the assets to either grow their own food or rely on others for charity. In Sen's own words:

"famines may be caused by a production failure, leading, (1.) to a direct decline of entitlements of those, such as peasants, whose means of survival depend on the food that they grow themselves; or (2.) to a sharp rise in the prices, thereby affecting the ability to command food on the part of those who have to buy food in the market" (Sen 1988:6).

Sen's analysis, therefore, disaggregates the reasons why a person or group may become vulnerable to hunger. This approach helps to highlight the difference between transitory or acute food insecurity and chronic hunger. Acute hunger could occur between harvests when either a) food supplies run low, or b) there is a lack of work, whereas chronic hunger would affect a population year round. Similarly, there is a difference between the starvation that occurs during a famine, when (perhaps) a weather event sparks a massive decline in food stocks, or when an economic problem reduces people's purchasing power,

and chronic hunger that occurs when an individual or family is unable to regularly purchase food (Chisholm and Tyers 1982).

Conclusions

Malthusians, neo-Malthusians and others who inherited his ideas, offer an explanation for the causes of poverty and hunger that are both simple and enduring. By focusing on quickly growing populations, the Malthusian arguments are easy to grasp and intuitively sound. When it comes to finding evidence for this theory, however, Malthus' ideas start showing their age. Fast growing populations, those that Malthus figured were heading into the abyss, consume the least amount of resources (in absolute terms as well as per capita). Therefore, from a global, national or regional perspective, these are not the populations who are having the largest impact on global environmental quality. Similarly, there is little correlation between famine, hunger, malnourishment or disease and population size, density or growth rates. The Malthusian pre-occupation with population, therefore, seems somewhat misguided, no matter how intuitively appealing. As a result, most scholars have expanded from a narrow focus on population to include the levels and types of consumption and technology in a society ($I=PAT$) as well as the fact that industrial activity can adversely affect remote areas (the ecological footprint and the shadow ecology).

An important concept of how many people planet Earth can support, which inspired some of the described approaches, is the notion of a carrying capacity. Notwithstanding the specific challenges of measuring carrying capacity, a consensus seems to be emerging that there must be limits to our activities and we need to observe these limits in order to be sustainable. Although Malthus underplayed our ability to innovate, there is no longer any question that

carrying capacity is influenced by human ingenuity, infrastructure, technology, educational resources, political institutions, international economic arrangements, land tenure systems, management skills and traditions, or the ability to mobilize in case of threats (Cohen 1995a).

Nevertheless, the Malthusian ideas had immense impact on drafting the population policies in the developing countries. The failure of these programs reflects inclination of the world powers to the “easy” solutions for underdevelopment.

In terms of food security, modern scholarship has more or less rejected Malthus’ preoccupation with population growth and now focuses on whether people have the ability to obtain food. This is an important shift because it moves from a deterministic model, where all regions with high populations were assumed to be suffering the same problems, and policies focused on birth control, to a less prescriptive approach that attempts to empower residents to find solutions that are tailored to local situations.

Selected References

- Chisholm, A. and R. Tyers. (1982) *Food Security: Theory, Policy and Perspectives From Asia and the Pacific Rim*. Toronto: Lexington Books.
- Cohen, J.E. (1995a) *How Many People Can the Earth Support?* New York: W.W. Norton.
- Cohen, J.E. (1995b) “Population Growth and Earth’s Human Carrying Capacity”, *Science*, 269, 341-346.
- Commoner, B. (1971) *The Closing Cycle*. New York: Knopf.
- Daily, G.C., and P.R. Ehrlich. (1992) “Population, Sustainability, and Earth’s Carrying Capacity”, *Bioscience*, 42.10, 761-771.
- Dauvergne. (1997) *Shadows in the forest: Japan and the Politics Of Timber in Southeast Asia (Politics, Science, and the Environment)*. Cambridge, MA: MIT Press.
- Davis, M. (2001) *Late Victorian Holocausts: El Nino Famines and the Making Of the Third World*. London: Verso.
- Dreze, J. and A. Sen. (1989) *Hunger and Public Action*. Oxford: Clarendon Press.
- Ely, R.T. and G.S. Wehrwein. (1948) *Land Economics*. New York: Macmillan Company.
- Ehrlich, A. and P.R. Ehrlich. (1970) *Population, Resources, Environment*. San Francisco, CA: W.H. Freeman.
- Escobar, A. (1995) *Encountering Development: the Making and Unmaking Of the Third World*. Princeton, NJ: Princeton University Press.
- Fischer-Kowalski, M. and C. Aman. (2001) Beyond IPAT and Kuznetscurve: Globalization As A Vital Factor Of Socio-Economic Factors Influencing Environmental Metabolism. *Population and Environment* 23, 7-47.
- Foster, P.B. (1992) *The World Food Problem*. London: Lynne Reinner Publ.
- Fraser, E. (2003) "Social Vulnerability and Ecological Fragility: Building Bridges Between Social and Natural Sciences Using the Irish Potato Famine as a Case Study", *Conservation Ecology*, 7, 1.
- Giljum, S. and K. Hubacek. (2001) *International Trade, Material Flows and Land Use: Developing A Physical Trade Balance for the European Union*. Interim Report. No. 01-059. International Institute for Applied Systems Analysis (IIASA), Laxenburg, A.
- Haddad, L.; J. Hoddinott and H. Alderman. (1997) *Intra-Household Resource Allocation in Developing Countries*. London: John Hopkins University Press.

- Hardin, G. (1974) "Lifeboat Ethics: the Case Against Helping the Poor", *Psychology Today*, 38-43, 124-126.
- Keopman, J. (1997) "The Hidden Roots Of the African Food Problem", in N. Visvanathan; L. Duggan; L. Nisonoff and N. Weigersma (Editors), *The Women, Gender and Development Reader*. London: Zed Books.
- Liebig, J.V. (1859) *Naturwissenschaftliche Briefe Über Die Moderne Landwirtschaft*. Leipzig & Heidelberg: Wintscher'sche Verlagsbuchhandlung.
- Malthus, T. (1976) *An Essay On Population*. New York: Norton Books.
- Meadows, D.H. and Club Of Rome. (1972) *The Limits To Growth. A Report for the Club Of Rome's Project On the Predicament Of Mankind*. Second Edition. New York: Signet.
- O'Grada, C. (1989) *The Great Irish Famine*. London: Macmillan.
- Paim, L. (1995) "Definitions and Measurements Of Well-Being", *Journal Of Economic and Social Measurement*, 21, 297-309.
- Rao, M. (No Date) *An Imagined Reality: Malthusiasm, Neo-Malthusiasm and Population Myth*. Cambridge, MA: Harvard Education.
- Ravallion, M. (1994) *Poverty Comparisons*. Paris: Harwood Academic.
- Rees, W.E., and M. Wackernagel. (1995) *Our Ecological Footprint: Reducing Human Impact on the Earth*. Gabriola Island: New Society Pub.
- Rostow, W.W. (1971) *Politics and Stages Of Growth*. Cambridge, UK: Cambridge University Press.
- Seidl, I., and C.A. Tisdell. (1999) "Carrying Capacity Reconsidered: From Malthus' Population Theory To Cultural Carrying Capacity", *Ecological Economics*, 31, 395-408.
- Sen, A. (1981) *Poverty and Famines*. Oxford: Clarendon Press.
- Sen, A. (1987) *Hunger and Entitlements: Research and Action*. Helsinki: World Institute for Development Economics Research (WIDER) United Nations University.
- Sen, A.K. (1988) "Food Entitlements and Economic Chains", in B. Lemay (Editor), *Science, Ethics and Food*. London.: Smithsonian Institute Press.
- Woodham-Smith, C. (1962) *The Great Hunger*. London: Penguin Books.
- York, R.; E.A. Rosa and T. Ditz. (2003) "STIRPAT, IPAT, and Impact: Analytic Tools for Unpacking the Driving forces Of Environmental Impacts", *Ecological Economics*, 46, 341-365.

Evan Fraser, Klaus Hubacek
Faculty of Environment, University of Leeds
Leeds. UK
E.D.G.Fraser@leeds.ac.uk
k.hubacek (at) leeds.ac.uk

Katarina Korytarova
Encyclopaedic Institute
Slovak Academy of Sciences
Bradacova 7, 851 02 Bratislava
Slovak Republic
encykory@savba.sk

Non-Profit Enterprises

Robert Scott Gassler

Introduction

Non-profit enterprises are those which are neither public nor profit-seeking. That is to say, the definition of a non-profit enterprise is that of an organization with no shareholders (Gassler 1986, Frumkin 2002). Such organizations include Greenpeace, the Catholic Church, the International Red Cross/Red Crescent, political parties, labor unions, trade associations, and consumer unions; and most revolutionary groups. Non-profit enterprises are also called voluntary, nonprofit, third-sector, charitable, civil-society, or nongovernmental organizations. Different writers make different distinctions among those terms. For example, in the study of development a distinction is made between indigenous organizations and international aid organizations. Nonetheless, they all suggest similar things about their structure, conduct, and performance.

Europeans (Defourny and Compos 1992) tend to link nonprofits and cooperatives in the “social economy”, but it is important to note that cooperatives exist to make a profit for their owners just as firms do; the difference is that in cooperatives the owners are the workers or customers. Economists also tend to classify nonprofits with public enterprises, but the latter are owned (at least partially) by governments and designed to seek profits. Nongovernmental organizations (NGOs), as they are usually called in international contexts, are to be distinguished from intergovernmental organizations, such as NATO or the United Nations, though their analysis in many ways may be similar. The same is true of households and families. Clubs are also usefully separated out as a distinctive category.

The economic analysis of non-profit enterprises began in the 1960s with the work of Ginzburg, Hiestand, and Reubens (1965), Boulding (1973, 1981) and Weisbrod (1975). Only the last of these really caught on and therefore only Weisbrod can really be said to be the founder of “nonprofit economics”. By the 1980s so much research had been done that it could be collected in edited volumes issued at frequent intervals: Clarkson (1980), White (1981), Rose-Ackerman (1986), James and Rose-Ackerman (1986), Powell (1987), Hodgkinson and Lyman (1989), James (1989), Gidron et al (1992), McCarthy, Hodgkinson, and Sumariwalla (1992), Hammack and Young (1993), and Anheier and Salamon (1998). Gassler (1990) provides a somewhat dated but useful survey, and Frumkin (2002) a useful but US-centered book-length introduction. Though the scholarship tends to be concentrated in the US, the scope of the research covers much of the world, and at the turn of the century a major comparative project was conducted by one of the field’s leading research centers (Salamon 1999).

Structure

The theory of governance of nongovernmental organizations is a special case of the theory of organizations in general, and in a way it could be seen as the general case and the others (governments, families, clubs, etc.) the special cases. Certainly theories of bureaucracy apply to any large organization, as anyone who has tried to telephone a US internet provider can testify. Non-profits have entrepreneurs, governing boards, employees, clients, governments, and other stakeholders. They are subject to government regulation, subsidy, and tax policy, and in turn many attempt to influence public policy in all fields.

In the early years of the field of non-profit studies, economists tended to assume that non-profits were headed by people whose

motives were no different from those of anyone else. Indeed the economists looked upon leaders of non-profits as pretending to be doing good while really just figuring out the right tax breaks and marketing schemes to do well. Those who studied the sector more closely however usually found a link between nonprofit behavior and altruistic motives.

The contemporary study of the economics of altruism dates from Boulding (1973) and Phelps (1975), but the founder and inspiration for many nonprofit economists must be considered David Collard, whose 1978 book provided a comprehensive demonstration that standard economic models of 'rationality' could easily incorporate altruistic motives. (Etzioni 1988 argues strongly that it cannot incorporate morality, but that is another matter.) Economists have been reluctant to disentangle the concepts of rationality and selfishness, but a few brave souls have tried (e.g., Margolis, 1982, Kinkor, 1992). The link between altruism and non-profits has thus been more tenuous than one might expect; nonetheless some progress has been made (Gassler, 1986; Rose-Ackerman, 1996).

Much of the literature has been devoted to one or another variation on the standard theme: zero profit in the long run (Austen-Smith and Jenkins 1985), or service versus revenue maximization (Steinberg 1986). Other points of view, such as the role of the entrepreneur (Young 1983), have not been neglected. The authors involved may or may not have acknowledged altruistic motives, couching their theories in sufficiently general terms that it may not have been necessary.

The relation between the entrepreneur (if any), the board, and the employees can be explored using various approaches which now go under the heading of the economics of organization. A number of approaches have been standard in economics for decades: principal-agent theory, public choice theory (but see Kelman 1987, Quiggin 1987,

Brannan and Pincus 1987) X-efficiency theory (Leibenstein 1966, 1979), and behavioral theory (Simon 1959, 1961, 1979). Other more heterodox approaches include institutionalism (Hodgson 1998) and radical political economy (Bowles and Gintis 1987). Each of these approaches has been touched on in the literature on non-profit enterprises (e.g. Chasse 1995 on institutionalism, Gunn 1997 on radicalism, and Jegers 2003 on agency theory), but there is much room for further research.

Conduct

Why are non-profit enterprises formed? Why is the non-profit form chosen over the profit-seeking one? One answer is provided by Weisbrod's famous 1975 article, which asserts that they are formed in response to problems of externalities, public goods & governance. For example, suppose that the government has decided, through majority rule, that a certain amount of some public good is to be produced. Majority rule, according to public choice theory, implies that the decisive vote belongs to the "median voter", i.e., the one whose views put him or her in the middle of the spectrum of opinion. To get a bare majority, it is necessary to secure all the votes to one side of the median voter plus that of the median voter as well. If however there are a substantial number of citizens who would have preferred more of the public good, then those people might form a non-profit enterprise to provide the rest. Weisbrod's paper includes a number of predictions, especially one that in relatively heterogeneous societies the nonprofit sector tends to be larger, a prediction that seems to be borne out by the facts (Salamon 1999).

Other approaches have also been followed. The earliest is 'contract-failure theory' (Hansmann 1980, Chillemi and Gui 1991), and the closely-related transaction-cost theory (Holtmann and Ullmann 1991, Krashinsky

1986), which has since become popular in economics in general. Rather than seeing non-profit status as a source of inefficiency, here it is seen as a drawing card. "If the quality of output is difficult to measure, and if contracts for future delivery are difficult to enforce, the nonprofit form may act as a signal assuring people that quality will not be sacrificed for private monetary gain." (Rose-Ackerman 1986:5).

What do non-profit enterprises do? Gassler (1986,2003) describes five sets of activities performed by non-profit enterprises (and governments) in an economy. First, they help create the environment in which economic activity takes place, by assisting in the formation of human tastes and preferences (social and cultural capital), and underwriting the basic research on which technological progress depends. Schools and universities figure prominently here.

Second, non-profit enterprises help provide the rules of the game for the economic system: developing customs concerning property rights laws & institutions, contract enforcement, and transaction costs. The educational system may very well make the difference between a society full of conflict and one in which disputes are settled amicably, without a thought to violence or lawsuits.

Third, non-profit enterprises assist in developing or implementing microeconomic policy toward the allocation of resources. Weisbrod's theory above shows how non-profits provide public goods. They also provide a different perspective in cases involving long-term time horizons or excessive risk. Politicians may be looking to the next election, CEOs to their retirement package, but those in non-profit enterprises have no such artificial cutoff to their time frame. Perhaps that is why in the US it is in the private universities, not the public ones, that we find schools of international affairs

and most of the programs concerned with world peace.

Fourth, non-profit enterprises help redistribute income, whether in cash or in kind, balancing the values of equity and efficiency. This is the area most people think of when they refer to "charitable institutions." Hospitals, nursing homes, shelters for the homeless, and religious charities are examples of this.

Fifth, they have limited roles in macroeconomic stabilization, mostly in encouraging or discouraging economic growth, depending on how they stand on the relation between, for example, economic growth and the environment.

To be sure, many organizations do not consciously pursue such lofty goals, but many do, and the others can sometimes be related indirectly. For example, trade associations may facilitate professionalism and transfer of economic information, both of which are beneficial, but they may also attempt to restrain trade, which is not.

The literature of social science is rife with references to "unintended consequences", usually negative ones. Frumkin (2002:ch2) however points out that the non-profit sector produces a very important byproduct from a societal point of view: it provides people with practice in democracy. The local associations and political groups that are included in the sector enable people to exercise leadership and express their values in a way that strengthens the political system (if it is democratic, and weakens it if it is dictatorial). This assertion is contested in the case of African countries by Allen (1997).

Performance

Early discussions of non-profit enterprises stressed the abuses that managers often committed in order to gain extra income for themselves (Etzioni and Doty 1976). They also stressed the inefficiencies that were

alleged to result directly from the non-profit status of the organization. For example, non-profit hospitals may have a bias toward higher quality care for a few rather than lower-quality care for the many. They may also fail to respond to the “market” in the same way as firms, when there is an increase in need or demand (Newhouse 1970).

More recently economists have begun to stress the possibility that non-profits may be more efficient than their profit-seeking counterparts. Weisbrod and Schlesinger (1986) found that by one measure non-profit nursing homes were more trustworthy than state-run homes, and profit-seeking homes were less. James (1989) claims that non-profits do respond to demand by filling in gaps left by the state: where public universities serve the elite, non-profits work to serve the poor; where the state universities are open to all, the non-profits may constitute the élite universities.

There is also progress in separating the effects of non-profit-ness *per se* from other causes of inefficiency. For example, suppose some aspects of output were measurable and some not. Paying employees according to the measurable outputs would lead them to neglect the other aspects of their job. This may be benign in the case of textile production, where very little is unmeasurable, but could be detrimental in the case of “publish-or-perish” higher education. Non-profit enterprises may arise to employ people who will pay attention to the unmeasurable aspects of their jobs out of professional pride or altruistic feelings, rather than neglect them for greater income. But this is equally true for those who work for profit-making firms (Frumkin 2002:ch3).

Gassler (1997, 2003) constructs a general model of organizations on standard economic principles and shows that profit and non-profit organizations differ little in their theoretical economic efficiency. He finds

however that the motives of the managers matter: non-profit enterprises should be headed by altruists, and profit-making firms by selfish people.

Practice

Clearly the activities of non-profit enterprises cannot be easily evaluated by profit criteria, but perhaps most are amenable to extended cost-benefit analysis. Since many people believe that a non-profit is less efficient than a profit-seeking firm, there is pressure to “run it like a business.” The commercialization of the non-profit sector has become an issue in the US.

The problem of commercialization is intertwined with another one: the relation between the non-profit sector and the government. In most Western countries, non-profit enterprises are heavily subsidized by the state. The reason often is that the government is considered better able to raise money, by taxation, but the non-profit enterprises, with less rigid rules, are considered better able to deliver services, especially those of a welfare state (Salamon 1999). This pattern existed in the US as well but was little recognized by the Reagan Administration in 1980 (Salamon and Abramson 1982; Palmer and Sawhill 1984). The administration cut social spending, under the impression that government spending substituted for non-profit spending. Non-profit revenues fell, until the organizations found ways to raise funds by selling things and otherwise commercializing. By the end of the decade questions were being raised about whether the non-profit sector should exist at all, since it so greatly resembled the business sector.

The question of whether non-profits should be run like businesses is both academic and practical. Programs to train non-profit managers have grown in the US and elsewhere, sometimes in business schools

as specialties within the MBA, sometimes in schools of public affairs, and in a few cases in separate programs of their own. Policy issues concerning non-profits include questions such as whether non-profits crowd out profit-making activity (Steinberg 1991) and whether the tax advantages given non-profits allow them to engage in “unfair” competition with profit-making firms.

In other countries, the issues differ. In Central and Eastern Europe, for example, the collapse of communism owes much to the work of non-profit organizations such as Solidarity in Poland and the Civic Forum in Czechoslovakia. Nonetheless, in some countries the very necessity for non-profit enterprises has been doubted—ironically by then-Prime Minister Vaclav Havel in the Czech Republic.

Conclusion

One thing is clear from the research so far: non-profit enterprises are complex phenomena and answers are not simple. Non-profit enterprises appear to perform significant tasks in a market economy, but what they do is unclear if looked at through standard theories. Some modifications are needed even to introduce them into a model of the economy. Several approaches have done just that.

So far much of the research into governance of non-profit enterprises is economic. There still needs to be more research by other social scientists into how non-profit enterprises work and how they fit into society.

Selected References

Allen, Chris. (1997) “Who Needs Civil Society?”, *Review of African Political Economy*, 73, 329-337.

Anheier, Helmut K. and Lester M. Salamon. (1998) *The Nonprofit Sector in the Developing World: A Comparative*

Analysis. Manchester: Manchester University Press.

Austen-Smith, David and Stephen Jenkins. (1985) “A Multiperiod Model of Nonprofit Enterprises”, *Scottish Journal of Political Economy*, 32, 2, 119-134.

Boulding, Kenneth E. (1973) *The Economy of Love and Fear: A Preface to Grants Economics*. Belmont, California: Wadsworth.

Boulding, Kenneth E. (1981) *A Preface to Grants Economics: The Economy of Love and Fear*. Second edition. New York: Praeger, (Praeger Studies in Grants Economics).

Bowles, Samuel and Herbert Gintis. (1987), *Democracy and Capitalism: Property, Community and the Contradictions of Modern Social Thought*. New York: Basic Books.

Brennan, Geoffrey and Jonathan Pincus. (1987) “Rational Actor Theory in Politics: A Critical Review of John Quiggin”, *Economic Record*, 63, 180, 22-32.

Chasse, J, Dennis. (1995) “Nonprofit Organizations and the Institutional Approach”, *Journal of Economic Issues*, 29, 2, 525-533.

Chillemi, Ottorino and Benedetto Gui. (1991) “Uninformed Customers and Nonprofit Organization: Modelling ‘Contract Failure’ Theory”, *Economics Letters*, 35, 5-8.

Clarkson, Kenneth W. and Donald L. Martin. (1980) (Editors) *The Economics of Nonproprietary Organizations*. Greenwich. Connecticut: JAI Press.

Collard, David. (1978) *Altruism and Economy: A Study in Non-Selfish Economics*. New York: Oxford.

Defourny, Jacques and Jose L. Monzon Compos. (1992) (Editors) *Economie sociale: Entre economie capitalist et economie publique, The Third Sector: Cooperative, Mutual and Nonprofit Organizations*. Brussels: De Boeck.

- Etzioni, Amitai and Pamela Doty. (1976) "Profit in Not-for-Profit Corporations: The Example of Health Care", *Political Science Quarterly*, 91, 3, 433-453.
- Etzioni, Amitai. (1988) *The Moral Dimension: Toward a New Economics*. New York: Free Press.
- Frumkin, Peter. (2002) *On Being Nonprofit: A Conceptual and Policy Primer*. Cambridge, Massachusetts: Harvard University Press.
- Gassler, Robert Scott. (1986) *The Economics of Nonprofit Enterprise: A Study in Applied Economic Theory*. Lanham, Maryland: University Press of America.
- Gassler, Robert Scott. (1990) "Nonprofit and Voluntary Sector Economics: A Critical Survey", *Nonprofit and Voluntary Sector Quarterly*, 19, 2, 137-149.
- Gassler, Robert Scott. (1997) "The Economics of the Nonprofit Motive: A Suggested Formulation of Objectives and Constraints for Firms and Nonprofit Enterprises", *Journal of Interdisciplinary Economics*, 8, 4, 265-280.
- Gassler, Robert Scott. (2003) *Beyond Profit and Self-Interest: Economics with a Broader Scope*. Cheltenham: Edward Elgar, forthcoming.
- Gidron, Benjamin, Ralph M. Kramer and Lester M. Salamon. (1992a) *Government and the Third Sector: Emerging Relationships in Welfare States*, San Francisco: Jossey-Bass (Jossey-Bass Nonprofit Sector Series and Jossey-Bass Public Administration Series).
- Ginzberg, Eli, Dale L. Hiestand and Beatrice G. Reubens. (1965) *The Pluralistic Economy*, New York: McGraw-Hill.
- Gunn, Christopher. (1997) "The Nonprofit Sector: Radical Potential?", *Review of Radical Political Economics*, 29, 3, 92-102.
- Hammack, David and Dennis R. Young. (1993) (Editors) *Nonprofit Organizations in a Market Economy: Understanding New Roles, Issues and Trends*. San Francisco: Jossey-Bass.
- Hansmann, Henry. (1980) "The Role of Nonprofit Enterprise", *Yale Law Journal*, 89, 5, 835-901.
- Hodgson, Geoffrey M. (1998) "The Approach of Institutional Economics", *Journal of Economic Literature*, 36, 1, 166-192.
- Hodgkinson, Virginia A. and R. W. Lyman. (1989) (Editors) *The Future of the Nonprofit Sector: Challenges and Policy Considerations*. San Francisco: Jossey-Bass.
- Holtmann, Alphonse G. and Steven G. Ullmann. (1991) "Transactions Costs, Uncertainty and Not for Profit Organizations: The Case of Nursing Homes", *Annals of Public and Cooperative Economy*, 62, 4, 641-54.
- James, Estelle and Susan Rose-Ackerman. (1986) *The Nonprofit Enterprise in Market Economies*. New York: Harwood.
- James, Estelle. (1989a) (Editor) *The Nonprofit Sector in International Perspective: Studies in Comparative Culture and Policy*, New York: Oxford. (Yale Studies on Nonprofit Organizations.)
- James, Estelle. (1989b) "The Private Nonprofit Provision of Education: A Theoretical Model and Application to Japan", Chapter 2 of James (1989a).
- Jegers, Marc. (2002) "The Economics of Non Profit Accounting and Auditing: Suggestions for a Research Agenda", *Annals of Public and Cooperative Economics*, 73, 3, 429-451.
- Kelman, Steven. (1987) "'Public Choice' and Public Spirit", *The Public Interest*, 87, Spring, 80-94.
- Kinkor, Jiří. (1992) "Ekonomie Altruismu", *Politická Ekonomie*, 6, 767-773.
- Krashinsky, Michael. (1986) "Transaction Costs and a Theory of the Nonprofit Organization", in Rose-Ackerman, 1986.

- Leibenstein, Harvey. (1966) "Allocative Efficiency v. X-Efficiency", *American Economic Review*, 56, 392-415.
- Leibenstein, Harvey. (1979) "A Branch of Economics is Missing: Micro-Micro Theory", *Journal of Economic Literature*, 17, 2, June, 477-502.
- Margolis, Howard. (1982) *Selfishness, Altruism and Rationality: A Theory of Social Choice*. Chicago: University of Chicago Press.
- McCarthy, Kathleen; A. Virginia Hodgkinson and Russy D. Sumariwalla. (1992) (Editors) *The Nonprofit Sector in the Global Community: Voices from Many Nations*. San Francisco: Jossey-Bass.
- Newhouse, Joseph P. (1970) "Toward a Theory of Non-Profit Institutions: An Economic Model of a Hospital", *American Economic Review*, 60, 64-74.
- Palmer, John R and Isabel Sawhill. (1984) (Editors) *The Reagan Record: An Assessment of America's Changing Domestic Priorities*. Boston: Ballinger.
- Phelps, Edmund S. (1975) (Editor) *Altruism, Morality and Economic Theory*. New York: Russell Sage.
- Powell, Walter. (1987) (Editor) *The Nonprofit Sector: A Research Handbook*. New Haven: Yale.
- Quiggin, John. (1987) "Egoistic Rationality and Public Choice: A Critical Review of Theory and Evidence", *Economic Record*, 62, March. 10-21.
- Rose-Ackerman, Susan. (1986) (Editor) *The Economics of Nonprofit Institutions: Studies in Structure and Policy*. New York: Oxford. (Yale Studies on Nonprofit Organizations.)
- Rose-Ackerman, Susan. (1996) "Altruism, Nonprofits and Economic Theory", *Journal of Economic Literature*, June, 34, 2, 701-728.
- Salamon, Lester M. and Alan J. Abramson. (1982) *The Federal Budget and the Nonprofit Sector*. Washington, D.C.: Urban Institute Press.
- Salamon, Lester M.; Helmut K. Anheier; Regina List; Stefan Toepler; S. Wojciech Sokolowski and Associates. (1999) *Global Civil Society: Dimensions of the Nonprofit Sector*. Baltimore: The Johns Hopkins Center for Civil Society Studies. (The Johns Hopkins Comparative Nonprofit Project.)
- Simon, Herbert A.. (1959) "Theories of Decision-Making in Economics and Behavioral Science", *American Economic Review*, 49, 3, June, 253-283.
- Simon, Herbert A. (1961) *Administrative Behavior*. New York: Macmillan.
- Simon, Herbert A.. (1979) "Rational Decision-Making in Business Organizations", *American Economic Review*, 69, 4, 493-513.
- Steinberg, Richard. (1986) "The Revealed Objective Functions of Nonprofit Firms", *Rand Journal of Economics*, 17, 4, Winter, 508-526.
- Steinberg, Richard. (1991) "Does Government Spending Crowd Out Donations? Interpreting the Evidence", *Annals of Public and Cooperative Economics*, 62, 4.
- Weisbrod, Burton A. (1975) "Toward a Theory of the Nonprofit Sector in a Three-Sector Economy", in Edmund S. Phelps (Editor). Revised version reprinted as Ch. 1 of Rose-Ackerman (1986), Revised and reprinted in Weisbrod (1988).
- Weisbrod, Burton A and Mark Schlesinger. (1986) "Public, Private, Nonprofit Ownership and the Response to Asymmetric Information: The Case of Nursing Homes", Ch. 7 of Rose-Ackerman, 1986, 133-151.
- Weisbrod, Burton A. (1988) *The Nonprofit Economy*. Cambridge, Massachusetts: Harvard.

White, Michelle D. (1981) (Editor) *Nonprofit Firms in a Three-Sector Economy*. Washington, D.C.: Urban Institute Press, 1981.

Young, Dennis. (1983) *If Not for Profit, For What? A Behavioral Theory of the Nonprofit Sector Based on Entrepreneurship*. Lexington, Massachusetts: Lexington Books.

Internet Sites

Global

Idealist.org. www.idealists.org

Europe

Institute for the Study of Civil Society
www.civitas.org.uk

Union of International Associations.
www.uia.org

European Foundation Centre. www.efc.be

USA

Center for Civil Society Studies.
www.jhu.edu/~ccss

Independent Sector.
www.independentsector.org

Foundation Centre. fdncenter.org

Institute for Nonprofit Organisation
Management. www.inom.org/

*Robert Scott Gassler
Vesalius College
Vrije Universiteit, Brussels
rsgassle@vub.ac.be*

Nuclear Energy

Jack Reardon

Introduction

Richard Heyman wrote that if all scientific knowledge was destroyed and we could only pass one sentence to the next generation it would be the atomic hypothesis, “that all things are made of atoms—little particles that move around in perpetual motion, attracting each other when they are a little distance apart, but repelling upon being squeezed into one another” (Feynman et al. 1963:1).

Leucippus, in the fifth century BC called these tiny, indestructible particles atoms, from the Greek ‘not to cut.’ The atomic hypothesis, first articulated by Leucippus has since guided scientist’s search for nature’s ultimate building blocks.

The pre-Socratic philosophers assumed only four elements—earth, air, fire and water; today however, we have identified 112 elements that by definition cannot be further decomposed. The atom is the smallest constituent component of an element. Slice an element in half, carbon for example, and we have two pieces of carbon. Keep slicing and we obtain smaller pieces of carbon until we eventually reach the carbon atom. If we slice any further, we no longer have the carbon atom but something different.

Contrary to Leucippus, the atom is not indestructible, but composed of electrons, protons and neutrons (which in turn are composed of even tinier sub-particles.) Electrons have a negative charge and populate shells surrounding the nucleus. The electrons in the outermost shell (the valence) determine the element’s bonding potential. The fewer electrons in the valence, the more reactive that atom is; and conversely, a full valence renders the atom inert. Protons and neutrons, with positive and neutral charges respectively, constitute the nucleus which

accounts for 99 percent of the atom’s mass but less than one percent of the volume—the atom is mostly empty space.

The uniqueness of an element is determined by the number of protons in the nucleus, which is also the atomic number of that element. Hydrogen, for example with one proton has atomic number 1; oxygen with eight protons has atomic number 8; and uranium, the heaviest natural element with 92 protons has atomic number 92.

A strong, nuclear force—the most powerful force in nature—operates at the subatomic level to keep the nucleus intact, otherwise the protons would repel each other and all matter would crumble. The strong force is 100 times greater than the electromagnetic force—which attracts opposite charges and repels like charges—and 10^{42} times greater than the force of gravity. The electromagnetic force operates on all protons in the nucleus; whereas the strong nuclear force operates only on nucleons (protons and neutrons) in its immediate vicinity. It is the tension between the electromagnetic and the strong force that makes nuclear reactions so powerful.

Albert Einstein hypothesized that mass and energy are interchangeable; that is, energy can be converted into mass and mass into energy. In his famous equation, $E=MC^2$, Einstein demonstrated that a little mass can produce a lot of energy since mass is multiplied by the square of the speed of light, equal to 186,000 miles per second.

Nuclear energy seeks to harness the energy inherent in the mass of the atom. One method, known as fission (from the Latin ‘to split’) involves hurling a neutron at an atomic nucleus to split the atom into two smaller nuclei. The energy released, which can be calculated according to Einstein’s equation, is far greater than fossil fuel combustion. The burning of one carbon atom releases four electron volts of energy, whereas the fission

of one atom releases 200 megaelectron volts (one million times an electron volt).

A second method of harnessing the atom's energy is fusion, which involves smashing two atoms together with enough force to overcome the strong nuclear force. Two hydrogen atoms pressed together, for example, will form helium (from the Greek sun god Helios) which is lighter than the constituent hydrogen atoms. The change in mass releases a large amount of energy, greater than the strong force and ten times greater than fission. Fusion requires extremely high temperatures to overcome the strong force (temperature is a measure of energy, therefore the higher the temperature the higher the energy). Such conditions exist in the sun's core with temperatures of 15 million degrees Celsius and pressure 340 billion times that of the earth at sea level. The energy released from the fusion of the sun's hydrogen nuclei heats and lights the Earth 90 million miles away.

As the number of protons increases from the lightest to heaviest elements, the electromagnetic force continuously increases; whereas the strong force per nucleon first increases, then stabilizes before decreasing, eventually reaching a point where the nucleus is unstable. The strong force per nucleon is most binding for iron (atomic number 26) and weakest for uranium. Uranium, discovered in 1789 and named after the planet Uranus, is therefore a logical fission candidate; and hydrogen with only one proton in the nucleus is the perfect fusion choice.

Heavy atoms emit particles from the nucleus in order to become lighter and thus more stable. This process was named radioactive decay by Pierre and Marie Curie in 1896. An interesting feature of radioactive decay is that in any given sample, it takes the same amount of time for half the sample to decay; hence the term half-life.

We know from Einstein's equation that

radiation particles have high energy—the radioactivity from uranium, for example, heats the earth's core and moves its tectonic plates. Three particles of radioactive decay have been identified: alpha rays, the heaviest and slowest, are comprised of 2 protons and 2 neutrons which is a helium nucleus; beta particles which are high-speed electrons; and gamma rays, electrically neutral and similar to X-rays, but with a much shorter wavelength. Since gamma rays are electrically neutral, they are not slowed by collisions with target materials; rather they are highly penetrating and can easily penetrate human skin.

Uranium is naturally radioactive with a half-life equal to the age of the earth. Uranium will eventually decay into lead, an element with half the mass of uranium and non-radioactive. Interestingly, lead, a heavy dense element effectively shields nuclear radiation.

The high energy of radioactive decay causes extensive damage to living cells. Radiation fractures proteins and nucleic acids, inhibiting their function and resulting in loss of cell vitality. Radiation can also rupture cell membranes, decrease enzyme activity and in some cases, initiate cancer.

Harnessing the Atom's Energy

In 1938 the German physicists Otto Hahn, Lise Meitner and Fritz Strassmann discovered that bombarding the nucleus of a uranium atom with neutrons could result in fission. If a neutron is hurled at a uranium nucleus, the strong nuclear force will absorb the neutron, creating the unstable isotope U-236, which is chemically identical to uranium except slightly heavier. (Isotopes have identical chemical properties but differ in the number of neutrons in the nucleus, thus their masses are slightly different.) The energy released from fissioning U-236 is calculated according to Einstein's equation.

In 1940 a team of British scientists discovered that hurling a neutron at a uranium nucleus could initiate a chain reaction. Specifically, one neutron will fission the nucleus into two smaller nuclei, releasing 2.5 neutrons, each of which can be hurled at another nuclei, releasing 2.5 neutrons, and so on. The fission process continues exponentially as long as there is fissile material.

The same year, plutonium, named after the planet Pluto, was discovered at the University of California Berkeley. Plutonium, with atomic number 94 is transuranic. All elements with an atomic number greater than uranium are transuranic and must be manufactured synthetically, although trace amounts of Plutonium and Neptunium (atomic number 93) are found in uranium ore. Transuranic elements are also radioactive with much shorter half-lives than uranium and more intense radiation. Plutonium is highly fissile and given its propensity to self-ignite, ideal for bombmaking. It is also used peacefully in radioisotopes and thermonuclear generators.

Albert Einstein exhorted President Roosevelt in 1939 to develop an atomic bomb in order to defeat Germany. The United States urgently undertook this task after December 7, 1941. Under the covert Manhattan Project, the US successfully tested a nuclear bomb on July 16, 1945. On August 4, an atomic bomb (made from 60kg of uranium) was dropped on Hiroshima and on August 9, a bomb (made from 8 kg of plutonium) was dropped on Nagasaki.

After the Second World War, attention turned to the peaceful development of nuclear energy. The first nuclear reactor to produce electricity was in Idaho, December 1951. The USSR built the world's first commercial nuclear generator in 1954. The British built a nuclear reactor in 1956, followed by France (1959) and Canada 1962.

Nuclear power currently generates 19

percent of the world's electricity, up from 1 percent in 1960. Seventeen nations rely on nuclear energy for at least 25 percent of their electricity, with Lithuania and France topping the list at 80 and 78 percent respectively. Today, thirty-one nations operate 440 commercial reactors (all nuclear fission) with a total of 366,821 Megawatts (International Atomic Energy Agency June 2005). Nuclear power remains a choice mainly for developed nations, with a typical reactor costing between four and seven billion dollars. The United States, France, Japan, Russia and the United Kingdom generate two-thirds of the total megawatts from nuclear power (IAEA 2005). The United States has 104 of the world's reactors followed by France (59) and Japan (53).

Of the 29 reactors currently under construction, eight are in India and three are in China. Central and South America have only four reactors (two each in Brazil and Argentina), the African continent has two, both in South Africa, while the Middle East has none, although one is currently under construction in Iran. The Czech Republic, Slovakia, Romania and Bulgaria have announced plans to build new nuclear plants in order to reduce reliance on Russia—a major supplier of natural gas to Eastern Europe (Sovich 2005).

Just about all nuclear energy is used for electricity, although some is used for medical and industrial purposes. Fifty-six nations operate 284 medical and industrial research reactors; and in addition, 150 ships are powered by nuclear energy.

In June 2005, a six-nation consortium (US, Russia, China, Japan, South Korea and the European Union) selected Cadarache in southern France to build a prototype fusion reactor. It is expected to begin operation in 2016 with an estimated cost of 10 billion Euros. The Consortium promises to build a commercial reactor in Japan. Realistically,

however, energy from fusion is decades away, perhaps by 2050.

International Regulatory Structure of Nuclear Power

Oil was used as an incendiary weapon by ancient civilizations and today wars are fought over oil. Although all types of energy have dual military and civilian purposes, only nuclear energy has the capacity to destroy civilization. How to promote peaceful uses of nuclear energy while preventing the spread of destructive nuclear weapons is a central policy concern.

In 1957, a United Nations mandate established the International Atomic Energy Agency (IAEA) to promote the peaceful use of nuclear energy while inhibiting its use for military purposes. Headquartered in Vienna, it is an intergovernmental forum for scientific and technological cooperation and provides international safeguards against misuse. It also is responsible for evaluating the safety of the world's nuclear reactors. In 1997, the IAEA established the Additional Protocol which gives inspectors greater rights, boosting the IAEA's ability to detect undeclared nuclear activity.

In 1968 the Nuclear Non-Proliferation Treaty was signed by the existing nuclear powers (US, USSR, England, France and China) and 139 non-nuclear nations. The Treaty, made permanent in 1995, currently has 189 signatories—testimony that most nations believe that nuclear proliferation can only contribute to destabilization. The objective of the Treaty is to limit nuclear weapons to the five nations possessing nuclear weapons at that time (not coincidentally the five permanent members of the UN Security Council). The non-nuclear states agree not to develop nuclear weapons, while the nuclear powers agree to liquidate their stockpiles. Under the Treaty, all signatory states have the right to peacefully

pursue nuclear power. Signatory nations are required to report nuclear material they possess and must agree to audits and inspections. In the 1990s, the IAEA began inspecting violations of the Nuclear Non-Proliferation Act.

Most of the world's nuclear weapons are owned by Russia and the United States. An Arms Reduction Agreement signed in 2002 by George Bush and Vladimir Putin called for the dismantling of tens of thousands of nuclear warheads by 2012.

Several shortcomings of the Treaty exist. One, the IAEA cannot enforce its terms; nor can any nation be forced to sign, although violations can be backed up by diplomatic, political and economic sanctions. Two, the IAEA has jurisdiction only over declared activity; yet undeclared activity beyond its purview can provide a ruse to develop nuclear weapons while ostensibly using uranium for peaceful purposes. A fine line separates military from civilian nuclear uses. It is speculated that North Korea and Iran developed nuclear weapons through this Treaty loophole. Finally non-signatory nations have no obligation and can develop nuclear weapons on their own, as did India, Pakistan and Israel. As a consequence of the high profile cases of Iran and North Korea, and the risk of terrorist acquisition of nuclear weapons, the consensus is that the Treaty needs to be revamped.

Major uranium exporting countries, such as Canada and Australia, cooperate with IAEA safeguards under the auspices of the Nuclear Non-Proliferation Treaty to ensure that uranium exports do not contribute to the proliferation of nuclear weapons. The 44 member Nuclear Suppliers Group coordinates the control of nuclear exports, refusing to sell enrichment and reprocessing equipment to any state not already possessing full-scale functioning equipment and technology.

The spread of nuclear nations increases the

likelihood of war and the possibility of a backdoor sale of nuclear weapons to terrorists. At the same time, the current geopolitical situation with one superpower encourages the development of nuclear weapons to assert national sovereignty. The Indian Foreign Minister, for example, asserted after India successfully tested a nuclear bomb in 1998, that there will no longer be “nuclear apartheid in the world.”

Nuclear Fuel Cycle

Uranium is found throughout the Earth's crust soil at an average concentration of two parts per million. In some parts of the world, uranium deposits are sufficiently high to justify mining. Canada produces 29 percent of the world's uranium, followed by Australia (21 percent) and Kazakhstan, Niger and Russia, each with 9 percent. Underground mining accounts for 41 percent of production, followed by open-pit (28 percent) and in-situ leaching (20 percent).

The mining of uranium ore is similar to metalliferous mining in that pits and shafts are dug, overburden and waste is removed and the land must be rehabilitated. The main difference is, of course, the danger of radioactivity, particularly gamma radiation in high grade ore. In addition, uranium mining releases radon, a naturally occurring gas found in most rocks and a decaying product of uranium. Radon can cause lung cancer in significant doses.

Uranium ore contains less than 0.1 percent uranium. The solid waste products from the milling operation are known as tailings and contain 85 percent of the original radioactivity and practically all of the radon. Tailings are placed in a water dam to attenuate radioactivity, then buried and covered with vegetation.

Disagreement exists over an acceptable exposure of radiation. The scientific consensus is that at low levels of

exposure—below 100m/sv per year (milliSievert; a measurement of radiation exposure, equivalent to joules per kilogram)—the human body's natural repair mechanism can quickly repair cellular damage. Natural radiation from rocks, soil, space and even the human body, ranges from 2msv/year to 50 msv/year in some parts of Europe, Iran and India. Uranium miner exposure to radioactivity is only 3 msv/year, slightly above the level of natural radiation, thanks to adequate ventilation and dust suppression. The probability of cancer increases with dosages above 100msv. A single dosage of 5000m/sv, for example, would kill half the recipients within one month.

Ninety-nine percent of the extracted ore is the non-fissile isotope U-238 and one percent is the fissile isotope U-235. Both uranium isotopes are radioactive: U-238 emits alpha radiation with a half-life equal to the age of the earth and U235 emits gamma rays with a half-life of 704 million years.

At the second stage of the fuel cycle, a three-step process enriches the ore to 4 percent U-235 to enable a fissile reaction. First, the uranium is oxidized to create U_3O_8 , then converted to the gas uraniumhexafluoride UF_6 . The heavier U-238 is separated from the lighter U-235. The gas is then reconverted to a solid: uranium oxide, pressed and baked at high temperatures and collected into rectangular fuel pellets. The main risk at this stage is radiation exposure to toxic chemicals, necessitating handling precautions similar to other chemicals. The enrichment plants also emit Chlorofluorocarbons (CFCs) which contribute to ozone depletion. It should be mentioned that both the extraction and enrichment stages are highly energy intensive, a fact overlooked by most nuclear advocates. Coal-fired generators typically power the process, releasing significant

greenhouse gases into the atmosphere.

At the next stage, the fuel pellets are transported via ship and/or rail to the reactor. In addition to the risks affecting hazardous material transport, is the risk of theft from terrorists. Uranium is compact and easy to transport; it is also fungible, rendering its national origin difficult to ascertain. Major uranium exporters have enacted bilateral security agreements to prevent terrorist organizations from absconding with fissile material or nuclear weapons.

At the reactor stage, the enriched uranium pellets are stacked into fuel rods, typically about 1 cm wide by 3.5 meters. Several hundred rods are bundled together and placed in the reactor core, made of reinforced steel and concrete. A fission chain reaction is initiated that is tightly moderated by a control rod and a coolant. The control rod, made of non-fissile material such as cadmium, is placed in the reactor to control the speed of the fissile reaction by absorbing neutrons. The control rod slows neutrons to the surrounding temperature of the gas to facilitate capture by the nucleus.

The coolant, usually water (sometimes liquid sodium, or helium gas is used) removes the heat generated by fission reactions. The coolant becomes radioactive and cannot leave the premise, so a series of pipes transfer the heated contaminated water to uncontaminated water. The contaminated water is continuously circulated within the nuclear plant; while the uncontaminated water is either released into a large body of water or transmitted through a steam tower to cool before release. The heat from the fission process converts water into steam which then drives a turbine connected to a generator to produce electricity.

At the reactor stage the risk is a core meltdown, whereby the uranium fuel rods liquify through the core reactor releasing radiation and contaminating the water supply.

A complete core meltdown occurred at Chernobyl on April 26, 1986, caused by operator inexperience and a poorly designed reactor. In addition, this particular Soviet model used a graphite (carbon) operator, which oxidized to form carbon monoxide causing the reactor to catch fire. Forty-two people died with an untold significant increase in cancer incidence and deaths, especially thyroid cancer caused by the radioactive release of iodine 131.

A partial meltdown occurred at the Three Mile Island Reactor outside Harrisburg Pennsylvania on March 28, 1979 when the fuel rods liquefied but did not breach the container system. No one was directly injured, although radiation was released into the air and into the Susquehanna River.

The fissile reaction is controlled with numerous safeguards including human operators to prevent a runaway explosion. These control and safety features add about thirty percent to the cost of the average reactor. A nuclear bomb, on the other hand, utilizes the same fission process without any control mechanisms. A nuclear bomb works by first detonating an ordinary explosive surrounding a pile of fissile material. The conventional explosive blows inward, crushing the fissile material, thereby increasing its energy and the chance that a stray neutron will strike a nucleus. A neutron gun is also fired to add more neutrons, increasing the chance of a chain reaction. When the heat/pressure becomes too intense the bomb rips apart the casing.

Reactors use only 4 percent U-235, whereas nuclear weapons require either 90 percent U-235 or Plutonium to increase the probability of a self-sustaining nuclear reaction. The IAEA safeguards the enrichment process and, fortunately, few countries possess the technical expertise to produce weapons grade Uranium.

Converting raw material to energy

generates waste and nuclear fission is no exception. Specifically, three types of waste are generated: High level—containing mostly fuel rods; Intermediate—containing the chemical sludges, metal fuel cladding and contaminated materials from reactor decommissioning; and low level—the rags, filter, gloves, clothing from reactors, hospitals and industry. High level waste accounts for three percent of the total nuclear waste but 95 percent of the radioactivity with half-lives in thousands of years; whereas low level waste comprises 90 percent of the waste volume and only one percent of the radioactivity.

Although nuclear waste receives a lot of attention, the total amount of nuclear waste is approximately one percent of industrial waste. Nuclear advocates claim future generations will find uses for nuclear waste; thus they prefer to keep the waste above ground. Critics, on the other hand argue it is unethical to enjoy the benefits of nuclear energy while bequeathing the risks to future generations.

No country currently has a long-term disposal site, although most nations intend to have underground storage facilities by 2010. Other long-term options for waste disposal include drilling beneath the ocean floor and shooting the waste into space.

The fuel in the highly radioactive rods is spent after 12-24 months and is either stored temporarily on-site in water for immediate cooling (and later in concrete casings) or reprocessed. A recent report found that spent nuclear fuel in US reactors is somewhat vulnerable to either a direct terrorist attack or theft by terrorists (National Academy of Sciences 2005). The report recommended a number of provisions to attenuate the risk, including reconfiguring the fuel rod assemblies and additional water spray cooling systems.

Ninety-five percent of the spent fuel rods is U-238; one percent is Pl 239; one percent is

U-235 that did not fissile; and 3 percent other fissile products. Reprocessing collects the five percent fissile products and separates them into their constituent components.

Although the fissile process is controlled, errant neutrons continuously strike and weaken the reactor. The average life expectancy of a nuclear reactor is about forty years, and the final step in the fuel cycle is decommissioning the reactor and disposing the high level waste.

The fusion fuel cycle utilizes the hydrogen isotopes deuterium and tritium since they fuse easier than hydrogen. Deuterium, also called heavy water, is chemically identical to water except that it has one extra neutron and thus a greater mass. Deuterium constitutes 0.015 percent of all water, but since water is abundant, the global supply of deuterium, ten million million tons, is practically unlimited. Tritium, is radioactive and made from Lithium, the lightest metal and plentiful within the earth's crust.

For the same amount of inputs, a fusion power plant produces ten times more energy than fission with at about 1000 times less waste. A 1000 megawatt coal plant in one year uses 9000 tons of coal and generates 30,000 tons of carbon dioxide waste, 600 tons of sulphur dioxide and 80 tons of nitric oxide. A 1000 nuclear fission reactor utilizes 147 pounds of uranium and generates 6.6 pounds of radioactive waste. A fusion plant uses 1 lb. of deuterium and 1.5 lbs of tritium and generates 4.0 pounds of helium waste and trace amounts of radioactivity (Princeton Plasma Physics Laboratory 2005). Another advantage of fusion, is that it is not possible to use fusion products for nuclear weapons.

The main problem with fusion is keeping and confining the hot gas. At extremely high temperatures the fuel is no longer a gas but a plasma whereby the atoms have become ionized – separated from their atomic nuclei and incapable of bonding. A magnetic field

confines the gas and prevents it from touching the reactor. For a fusion reaction to be viable, it must produce more tritium than it consumes. This could be accomplished by coating the reactor lining with a blanket of lithium three feet thick; as the neutrons are thrown off, they strike the lithium and produce tritium. However, fusion requires far more energy than fission to produce the hot temperatures and keep the plasma self-sustaining. The technique of cold fusion could potentially obviate the need for high temperatures and thus high energy, but claims of initiating cold fusion have not been replicated.

Advantages of Nuclear Power

In a widely quoted 1954 speech, Lewis L. Strauss, the chair of the US Atomic Energy Commission boasted, "Our children will enjoy in their homes electrical energy too cheap to meter."

Although Strauss does not mention nuclear power, many assumed during the 1950s that nuclear power would play a vital role in advancing living standards. However, thanks to cheap oil prices and well-publicized accidents, nuclear energy failed to reach its potential.

Currently, however, nuclear energy has gained momentum thanks to global warming. The scientific consensus is that human activity, especially fossil fuel combustion contributes to global warming, an idea that has now become mainstreamed (Colvin 2005). Overall nuclear power, as a substitute for coal in the production of electricity, reduces carbon dioxide emissions by 2 billion tons per year. This is touted as the main advantage of nuclear power.

A second advantage is that uranium is highly concentrated. One atom of fissionable uranium produces 10 million times more energy than a single carbon atom. Stated a little differently, one ton of uranium produces

45 million kilowatt hours of electricity; whereas the same amount of electricity requires 20,000 tons of coal and 30 million cubic meters of natural gas. In addition, nuclear fission energy consumes less than ten percent of the energy it eventually creates, far better than fossil fuels.

Third, uranium reserves are abundant vis-à-vis fossil fuels. The consensus is that oil and natural gas supplies will peak sometime before the year 2025, causing prices to escalate. At present rates of consumption, fifty years of high grade uranium reserves exist, but since fuel is a minor cost of fission power, more expensive grades of uranium could be used with little affect on price. If the price of U_3O_8 were to double for example, the cost of fission power would increase by 30 percent and the electricity cost by about 7 percent; whereas doubling natural gas prices would increase the cost of electricity by 70 percent. Furthermore, reserves of U-238 are far more plentiful; it is estimated that anywhere from 10,000 to 5 billion years of U-238 exist.

Fourth, nuclear energy can reduce dependence on fossil fuels. Japan, for example, deficient of fossil fuels, has doubled its nuclear capacity since 1992 and relies on nuclear energy for 42 percent of its electricity. France dramatically increased its nuclear capacity after the OPEC embargo citing energy security and maintenance of living standards. Both the EU and the United States have cited energy security as a reason for a sustained nuclear commitment (World Nuclear Association June 2005).

Fifth, the nuclear industry has some 10,000 reactor years with an accident rate of less than 1 percent. This compares favorably with the fossil fuels especially the coal industry in which some 1000 miners die annually from occupational injury and thousands more suffer from debilitating disease. Although the risk of a catastrophic accident is ever-present,

nuclear energy advocates reminds us that the only core meltdown was in a poorly constructed and outdated Soviet era model.

Sixth, although renewable sources will play a prominent role in the future, at present only nuclear power can satisfy the demand for large-scale, highly intensive and reliable electricity (Huber and Mills 2005). Renewables, on the other hand, remain erratic and intermittent and are more poised to meet low intensity, dispersed needs.

Finally, development of nuclear energy stamps a nation with the imprimatur of modernity: possessing nuclear energy is a point of national pride, signaling that the nation possesses the intellectual sophistication of the modern age.

Arguments Against Nuclear Energy

While no form of energy is risk free (Andolouse 1997), the perceived dangers of nuclear power continues to galvanize opposition. The industry's most vexing problem is the failure to devise an adequate waste disposal plan, which is "easily the most lasting insignia of the twentieth century and the longest lien on the future that any generation of humanity has yet imposed. (McNeil 2000:313). Perhaps if an adequate waste disposal plan was implemented before commercial reactors were developed, public attitudes today might be less strident.

Second, is the threat of a radiation leak and/or a nuclear meltdown. The nuclear industry can rightfully boast of its low accident rate compared to fossil fuels; yet the risk remains, however small, of a cataclysmic accident affecting millions of people, since "existing reactors have not reached and will never reach a nuclear nirvana where catastrophes cannot happen" (Union of Concerned Scientists 2004:3). In addition, high profile accidents such as the 1999 radiation leak in Tokyo in which two people were killed and an August 2004 explosion in

Tokyo in which 4 people were killed, continues to galvanize the world's attention to the perils of nuclear energy.

Third, a full analysis of the nuclear cycle reveals its early stages are energy intensive. Large amounts of energy are needed, primarily from fossil fuels to build the reactor, mine and enrich the uranium, sequester and transport the waste and dismantle the plant (van Leeuwen and Smith 2004). This is an important point overlooked by nuclear advocates. Whether nuclear energy or natural gas emits more CO₂ depends on the grade of uranium ore: For rich ores, (at least 1 percent uranium content) the nuclear plant over its life cycle will emit 30 percent less CO₂; whereas using ore with less than 1 percent uranium results in more CO₂ emission. Unfortunately, the supply of rich ore, based on current rates of consumption, will last only fifty years (van Leeuwen and Smith 2004:6). Accordingly, Van Leeuwen and Smith conclude, "Nuclear power is not a viable way to substantially reduce CO₂ emission. It is no exaggeration to say that nuclear power can only exist because it is fueled by fossil fuels" (2004:3).

Fourth, although the IAEA has implemented safeguards, covert nuclear activities can be conducted under the guise of peaceful nuclear energy utilization. Since 2000, North Korea and Iran have developed nuclear weapons in such a way, and it is a safe bet that more nations will develop nuclear weapons, increasing the chance of war and of nuclear technology falling into the hands of terrorists. It should be pointed out, however, that the only use of nuclear weapons came when one nation—the United States—had a monopoly. And furthermore, eliminating nuclear weapons will not attenuate the risk of war.

Fifth is the risk of terrorists obtaining either the fuel, the technology or nuclear weapons. The Union of Concerned Scientists

brands this the gravest of all the threats facing the world. Although most Western facilities are heavily guarded, many in the former Soviet Union are not, increasing the probability of theft.

Sixth, nuclear energy is heavily subsidized: no commercial reactors are profitable on their own. Why should the public support energy programs that could lead to destructive weapons? On the other hand, fossil fuels are heavily subsidized: greenhouse gases are dumped into the atmosphere and the US military safeguards Middle Eastern oil.

Future of Nuclear Energy

Electricity is the fastest and most compact form of energy; it is necessary to activate microwaves, lasers, X-Rays, magnetic pulses, silicon chips, magnetic resonance imaging and highspeed wireless and much more (Huber and Mills 2005:16). Electricity demand, which accounts for 40 percent of energy demand (transportation and heat each account for 30 percent) is expected to double by 2025, increasing faster than total energy demand. Much of this growth will occur in developing nations where the demand for electricity is expected to increase 3.5 percent annually, compared to 1.6 percent for developed nations.

The demand for nuclear energy, however, is expected to decline to 11.8 percent of total energy by 2025 largely at the expense of natural gas (Energy Information Administration 2005). It is expected that many developed nations will let operating licenses for individual reactors expire due to substantial political opposition. Germany, for example, has pledged to become nuclear free by 2021, thereby closing 19 nuclear plants. The demand for nuclear energy will also be a function of developing technology. Hybrid vehicles that use electricity and fossil fuels can power automobiles and heat homes, thus bridging the chasm between the different

segments of energy demand.

Without a universally accepted ethical consensus to guide fuel choice, selection will be a function of relative price and perceived risk. If one fuel becomes scarce or associated risks increase, the demand will increase for competitive fuels. Barring any catastrophic accident, global demand for nuclear energy will continue to increase. It also must be recognized, however, that nuclear energy is not a panacea for global warming which remains the most pressing concern.

Selected References

- Audouze, Jean. (1997) *The Ethics of Energy*. United Nations Educational, Scientific and Cultural Organization. www.unesco.org
- Bennhold, Katrin. (2006) *Nuclear Energy is Making a Global Comeback*. energybulletin.net
- Collier, John G. and Geoffrey F. Hewitt. (2000) *Introduction to Nuclear Power*. Second Edition. New York: Taylor and Francis.
- Colvin, Geoffrey. (2005) "Nuclear Power is Back—Not a Moment Too Soon", *Fortune*. May 30. www.fortune.com
- Energy Information Administration/ Department of US Energy. (2006) *International Energy Outlook*. www.eia.doe.gov
- European Nuclear Society. (2008) "What is a Nuclear Reactor?" euronuclear.org/info/
- Feynman, Richard, Robert Leight and Matthew Sands. (1963) *The Feynman Lectures, Vol. 1*. Reading Mass.: Addison Wesley.
- Huber Peter and Mills, Mark Mills.(2005) *The Bottomless Well*. New York: Basic Books.
- International Atomic Energy Association. (2005) *Nuclear Power Plants Information*. June. www.iaea.org/
- Makhijani, Arjun and Scott Saleska. (1999) *The Nuclear Power Deception*. New York:

- Apex Press.
- McNeil, John. (2000) *Something New Under The Sun: An Environmental History of the Twentieth Century World*. New York: Norton.
- National Academy of Sciences. (2005) “Spent Fuel Stored in Pools at Some US Nuclear Power Plants Potentially at Risk from Terrorist Attacks; Prompt Measures Needed to Reduce Vulnerability.” April 26. National-Academies.org/
- Nuclear Suppliers Group. www.nuclearsuppliersgroup.org/
- Percebois, J. (2003) “The Peaceful Uses of Nuclear Energy”, *Energy Policy*, Volume 31, January, pp. 101-08.
- Princeton Plasma Physics Lab. “The Advantages of Fusion.” www.pppl.gov/fusion_basics/
- Sagan, Carl. (1980) *Cosmos*. New York: Random House.
- Sovich, Nia. (2005) “Europe’s New Nuclear Standoff”, *Wall Street Journal*, June 29, p. A13.
- Union of Concerned Scientists. (2004) “US Nuclear Plants in the 21st Century.” May. www.ucsusa.org/clean_energy/nuclear_safety/
- Van Leeuwen, Jan William Storm and Philip Smith. (2004) *Can Nuclear Power Provide Energy for the Future; Would it Solve the CO₂ Emission?”* www.stormsmith.nl
- World Nuclear Association. (2003) *Safety of Nuclear Reactors*. November. world-nuclear.org/info/
- World Nuclear Association. (2005) *Sustainable Energy*. June. www.world-nuclear.org/info/

*Jack Reardon
Department of Economics
Hamline University
Minnesota, USA*

Open Source Software Policy

Alan Isaac

Introduction

Few business, governmental, or scientific functions are untouched by software innovation. Software innovation also provides important support to creative endeavors in art and industry, as well as being a creative endeavor in its own right. Software innovation has become an important influence on our standard of living, and policy makers at the local, national, and international level express concerns about a “digital divide” between those who have access to the new technologies and those who do not. The success of free and open source software introduces the surprising possibility that the goal of promoting software access need not conflict with the goal of sustaining the rate of innovation, but this enticing possibility remains controversial. The appropriate policy stance for government to take toward free and open software is correspondingly controversial.

Free and Open Development

Under free and open development, developers produce technological innovation as a public good rather than as proprietary intellectual property. The “freedom” is primarily a matter of licensing: the developer licenses all persons or groups to use, freely redistribute, and openly modify the technology. The “openness” is primarily a matter of enablement: the developer discloses the technology in a way that unambiguously enables others to implement and modify it. Free and open (FO) development generates FO technology. Technology placed in the public domain, such as the TCP/IP protocols for internet communication and their public domain implementations, is also considered FO in this sense.

Free and open source software (FOSS) is a vibrant example of FO technology. FOSS is usually generated by FO development, but sometimes closed-source, proprietary software is rereleased as FOSS. FOSS is any software that has FO licensing and enablement characteristics (OSI 2004 provides an extended discussion of these characteristics.) Software that is not FOSS is usually called ‘proprietary’.

Since unambiguous enablement is a crucial component of FO development, and since only the source code is truly enabling for complex software applications, software development is considered to be free and open only if the source code is readily available and freely redistributable. (The source code is the human-readable, complete implementation of the software that programmers generally work with.) FOSS development does not require gratis distribution of source code, but in practice FOSS software has been usually provided on the internet for download without charge. Occasionally commercial software developers also provide relatively open access to their proprietary source code—usually in the presence of licensing fees enforced by patent or copyright claims—but in practice proprietary software is usually closed-source. Commercial software developers can nevertheless be involved in FOSS development. For example, a consortium of large firms recently has made substantial contributions to the development of the GNU/Linux operating system.

FO technology generally is not simply placed in the public domain. For example, most FOSS is copyrighted or patented, and trademarks play an important role as well. So contrary to the allegations of some observers, FOSS developers do not contest the importance of intellectual property rights: they assert them vigorously. This is particularly evident in the arena of licensing.

FO technology is often licensed under economically important restrictions. The most important example of a licensing restriction compatible with free and open development is that distributed modifications of a FOSS technology remain free and open. The most famous such licensing provisions are found in the GPL, which is one of the oldest and most popular FOSS licences. Proponents refer to software licensed under the GPL as “free software”, the idea being that it cannot be captured (i.e., rendered proprietary) by commercial interests using embrace-and-extend tactics.

In the arena of intellectual property rights, FOSS is more accountable than closed-source software products. For example, in the absence of whistle-blowers, violations of the GPL by commercial vendors are likely to go undetected. In contrast, source code availability implies that any copyright or patent violations in FOSS software will be easily discovered. It is however significant that FOSS developers use intellectual property rights differently than commercial developers. FOSS developers use existing intellectual property institutions to promote the immediate and sustained access of all other developers to their innovations, while commercial developers use the same institutions to restrict the access of other developers to their innovations. It remains unclear which approach is more likely to stimulate long-run software innovation, and the answer may well vary by product niche.

FOSS in Context

Early software development took place in an environment influenced by academic standards emphasizing the free exchange of knowledge. FOSS development emerged fairly naturally in this setting (Hippel and Krogh 2003). By the end of the 1970s, however, the tremendous economic value of computer software was evident, and software

subsequently became increasingly proprietary. Closed-source, proprietary software became a growth industry generating tremendous economic value—estimated at over US\$300 billion in 2003—and intellectual property rights in software were aggressively asserted (UNCTAD 2003). Initially firms relied heavily on trade secrets and copyright, since software was not considered patentable. However, by the end of the 20th century software patents were well established—if still controversial—and were experiencing explosive growth. This software patent explosion took place both in the U.S., where early caution about software patentability was largely discarded by the courts and by the US Patent and Trademark Office, and in Europe, where software patents are readily granted by the European Patent Office despite public controversy over the extent to which current law allows them.

It is natural to expect that FOSS development would become radically marginalized as software development shifted toward traditional business models. Instead something startling happened: despite the apparent inability of FOSS developers to directly derive financial rewards from their development efforts, FOSS development continues to flourish. We will refer to this outcome as ‘the open-source phenomenon’. The open-source phenomenon is economically important in two related ways: it is generating substantial economic value, and it has challenged policy makers to reexamine the conditions under which innovation can thrive.

Valuing FOSS

No credible estimate of the value of FOSS development and distribution is available, and any attempt to produce one will be fraught with difficulties. Attempts to value FOSS cannot use measures of value added as determined in competitive markets: most

FOSS production and distribution is not market activity. Neither can valuation attempts be based on the values of inputs used in software development: aside from the problematic relationship between production costs and product values, FOSS inputs are largely unknown. So the value of FOSS production is generally illustrated heuristically.

For example, around two-thirds of publicly accessible websites use the Apache web server, more than three times as many as use the most popular commercial web server. Google alone runs GNU/Linux on over 100,000 servers, implying almost all internet users are unknowing GNU/Linux users. The pervasiveness of FOSS is often overlooked because it remains relatively rare on the end-user desktop, but FOSS is making inroads even there: GNU/Linux is estimated to be on 15% of new computers sold in Western Europe and to have overtaken Apple in terms of market share on the desktop in the US. Use of the GNU/Linux operating system, already in widespread use for server applications, is approaching the desktop with enough speed that Microsoft considers it a major threat to sales of desktop Windows operating systems.

Many other FOSS applications are in wide use and are considered highly competitive with commercial products. Oft cited examples include the MySQL relational database, the sendmail mail transfer agent, the Mozilla web browser, and the interpreters for the Perl and Python programming languages. Also available as FOSS are excellent text editors, word processors, email clients, spreadsheets, presentation packages, games, and endless specialized utilities. If valued in terms of commercial substitutes, FOSS software adoption may be valued at tens of billions of dollars a year. Most observers agree that FOSS projects demonstrate unambiguously that high quality, commercially important, and very innovative development can take

place in the apparent absence of revenue-generating intellectual property rights.

FOSS is a Public Good

Economists characterize some goods as “public goods” based on technical and institutional characteristics of the good. Most commonly, economists define as a good to be a “public good” if it displays “non-rivalry” and “non-excludability”. A good displays non-rivalry to the extent that additional consumers do not reduce the value per consumer of existing production. A good displays non-excludability to the extent that anyone can consume it without payment. Imagine a good that is completely non-rivalrous and non-excludable: we call this ideal type a “pure” public good. With very minor qualifications, FOSS may be considered a pure public good.

Adequate production of a non-excludable good is often considered to require government action. Since individual consumers can decide that it pays to be a free rider (i.e., to consume the good without contributing toward it), markets do not force consumers to reveal their relative valuation of the good. Interesting points of comparison are television broadcasts and basic research, two classic examples of public goods.

Television broadcasts and the results of basic research are non-rivalrous in consumption for reasons that are primarily technological. However they are non-excludable for reasons that are primarily institutional. For example, patent law generally excludes from patentability the results of basic research. Television broadcasts highlight institutional considerations, since shifting institutions have rendered much television programming excludable. Even an open broadcast excludes those who do not have access to the right kind of receiver and monitor. A broadcast may be encrypted, excluding those who do not have

access to decryption technology. Watching a broadcast may be illegal, excluding those unwilling to break the law. This illustrates that most non-rivalrous goods are not inherently public or private: they are categorized by our institutions of exclusion.

Note an important stock/flow distinction: roughly speaking, a television broadcast provides a flow of transient consumption opportunities, while basic research adds to a stock of knowledge. Like basic research, FOSS development increases the outstanding stock of public knowledge about software production. Patented commercial research may also have this characteristic: in principal, patents should only be granted when enabling disclosure is provided. The difference between commercial research and basic research is largely institutional: society provides mechanisms for some individuals to exclude—at least temporarily—other individuals from freely using commercial research results. Most software development takes place within such institutions of exclusion, which allow private capture of the value of non-rivalrous goods. FOSS production takes place within different institutions, which are intended to limit or eliminate exclusion.

This highlights an important question: why does society develop such institutions of exclusion? This question has always been pressing for economists, who have developed detailed theoretical arguments that goods should be priced at their marginal cost of production. In the case of knowledge goods, the *raison d'être* of institutions of exclusion is to limit competitive pressures that could enforce marginal cost pricing. The standard pragmatic justification of institutions of exclusion is that the development of knowledge goods involves sunk costs, and so (it is claimed) these goods will be undersupplied unless institutional arrangements lure innovators with the

expectation of profits. The open-source phenomenon empirically calls this pragmatic justification into question, at least in the case of software development, and the example of basic research reminds us that society has already implemented alternative institutional solutions to the problems of motivating innovation in the presence of sunk costs.

Internationally, institutions of exclusion are still evolving. Even the legality of pure software patents is still in play internationally. The early 21st century will determine which international institutions of exclusion will be invoked for software innovations. The implications are far reaching. For example, a spreading acceptance of pure software patents may pose serious threats to FOSS development—and to innovative small firms who lack large defensive patent portfolios (Shapiro 2000).

FOSS and Innovation

Some researchers argue that FOSS is imitative rather than innovative, living parasitically off of innovations in the commercial sector. Whether FOSS or commercial, software innovation is highly imitative and incremental, so this will not be an easy argument to settle definitively. However most researchers consider FOSS highly innovative, citing major contemporary FOSS applications as well as the early history of software development, which was undeniably innovative and substantially free and open. Lerner and Tirole (2002) and Raymond (1998) emphasize that FOSS innovations have not just been in the products themselves but in the development process. High profile FOSS developments took place in innovative collaborative organizational structures, where technically sophisticated users often provided important impetus and even contributions for incremental FOSS innovation. Some researchers go so far as to claim that the most important innovation

produced by FOSS development was organizational (Weber 2004).

The traditional understanding of incentives to innovate is based in institutions of exclusion, which manifest in the form of intellectual property rights. For example, a patent holder is entitled to use legal institutions to exclude others from using the patented innovation. Patent systems are intended to increase the anticipated future profit flows from innovative activity, in the expectation that innovation will be undertaken in response. Optimal patent system design is an area of active research, and economists recognize many ways in which patents can encourage or stifle innovation. Nevertheless, policy makers and well-heeled industry representatives tend to assert without qualification that strong patent laws stimulate innovation and that the revenues secured by strong intellectual property rights are a necessary condition for growth.

The open-source phenomenon has therefore inspired substantial controversy. The voluntary creation of substantial economic value apparently outside “the market” has particularly provoked economists, who tend to see prices as a necessary conduit of the information that can allow efficient resource allocation. The open-source phenomenon poses troublesome questions to social scientists. Why does an abandonment of traditional economic rewards not forestall the allocation of factor inputs to the production of FOSS? Why does an inability to appropriate the economic value in intellectual property not stifle FOSS innovation? Why is a price mechanism not a necessary concomitant to these economic activities which are creating substantial economic value? By innovating and producing substantial economic value, FOSS presents social scientists and policy makers

with an important challenge to their understanding of the roots of innovation.

FOSS Developers

FOSS development has been primarily a private sector initiative. FOSS developers include private individuals, standards organizations, and commercial firms of diverse sizes. There is a corresponding diversity of motivations and governance structures. Individuals participate for many reasons, including love of programming challenges, desire to make a gift to future generations, ideological commitments to “free” software, expectation of increased personal productivity through incremental improvement of an actively used FOSS application, desire to achieve recognition in the FOSS community, and pursuit of remunerative skills or reputation (Lerner and Tirole; Weber 2004). The FOSS community appears to offer individuals unusual opportunities for symbiosis between self-regarding and altruistic ends. Commercial firms can also benefit from FOSS development in many ways, including the achievements of strategic advantages in oligopolistic product markets (such as desktop operating systems), the development of “absorptive capacity” by maintaining staff at the cutting edge of certain technologies (Cohen and Levinthal 1989), and the provision of complementary goods and services.

FOSS participants often stress the importance of altruistic and intrinsic motivations to their participation. When analyzing the FOSS community, economists have tended to discount these motivations, stressing instead the role of career signaling and peer recognition effects. For example, in an otherwise excellent paper, Lerner and Tirole 2002 seem eager to link ex post outcomes to possible ex ante motivations, with little attention to selection

bias. A positive result of this strategy has been the identification of many possible ways that individuals may profit by participation in FOSS development. For example, it may be easier to signal one's contribution to an open source project than to a proprietary project because the open source projects are structured to offer greater visibility to all participants. As Isaac and Park (2004) note, however, this does not explain why the innovator would eschew the assertion of property rights in the innovation (unless the GPL is involved). However, if one programmer's contribution leads others to invest in the project, his private valuation could increase (due to network effects) or costs decrease (due to a productivity effect owing to a higher stock of solved problems). While such explorations are important, it is also important that altruistic, artistic, and other intrinsic motivations appear important for some individual FOSS developers (Weber 2004).

Intrinsic motivations are unlikely explain the substantial involvement of commercial firms, which are increasingly important players in FOSS development. For example, the development the enterprise capabilities of GNU/Linux by the Open Source Development Lab has investment backing from Computer Associates, Fujitsu, Hitachi, HP, IBM, Intel, and NEC (among others). IBM in particular has been strongly backing GNU/Linux as a venue for selling services, applications, and hardware. There may also be strategic considerations: firms may expect a long-run benefit from commodification of the computer operating system, either directly (through lower costs) or indirectly (by promoting interoperability standards). Another long-run consideration concerns assurances of software availability and support: bankruptcy of a commercial vendor may end the support for a product, or increased license fees may end the

affordability of a product. Source code availability provides insurance against product discontinuance and licensing changes, and participation in FOSS development can help maintain the internal capacity to modify FOSS software when needed. Such strategic and long-term considerations are relevant to government as well.

Government Policy Toward FOSS

Government has numerous interests in policies that affect the use and development of FOSS, and this section highlights a few. Government use of FOSS has budgetary effects, which can be complex. FOSS is not directly taxed, but it may stimulate the growth of taxable industries. The absence of licensing fees directly reduces fiscal expenditures, but the full expenditure effect is determined by the total costs of operation under different configurations of applications and operating systems. Even if the total cost of operating in a FOSS environment proves higher—which appears unlikely, but debate on this issue continues—government may justify these costs as achieving other goals, including reduced vendor dependence, support for the local programming community, greater absorptive capacity, increased security, promotion of standards compliance, and greater transparency. When government believes future modifications will be an important aspect of its software use, FOSS may provide valuable independence from any single vendor (Bessen 2002).

In a country where software development is active, FOSS policies may affect the rate of innovation. It remains uncertain whether FOSS will accelerate or hinder software innovation in the long run, and the effects on innovation in software using industries are essentially unexplored. Countries lacking an active software industry may find that the promotion of FOSS builds absorptive

capacity and improves the trade balance, especially as technological and institutional pressures lead to greater enforcement of commercial software licensing. Security considerations may also lead to a reliance on open source: the debate over the relative security of FOSS and proprietary software remains active, but some governments have decided that there are advantages in being able to view the source code of imported software. Related to this, increasing international reliance on electronic voting has naturally increased interest in maintaining election transparency through the use of open source voting software.

Government Support of FOSS Development

To date, direct government support and subsidy of FOSS development has been rare. Government policy toward FOSS is rapidly evolving, however. Since FOSS development produces a public good with positive network externalities, direct subsidies might receive traditional economic justifications. However FOSS development can directly compete with profit oriented development, so the appropriate level of government support for FOSS development remains an active area of debate (Hahn 2002).

Recently some governments have taken steps to actively favor FOSS development. Is such government support is necessary to sustain FOSS development? The obvious answer is “yes”, just as it is necessary for commercial software development. Government is part of the institutional structure that is a crucial determinant of economic activity. A narrower question is more interesting: given the current institutional environment, are special subsidies or legal preferences necessary to sustain FOSS development? For a variety of reasons, turning empirically on the contemporary vitality of FOSS development,

economists tend to answer “no” (Bessen 2002, Evans 2002, Smith 2002). The most pressing core policy question is perhaps even narrower: given the current institutional environment, would the development of special subsidies or legal preferences for FOSS prove socially beneficial? Here the understanding of social benefit should be broad, including everything from stimulating software innovation to reducing dependence on foreign vendors to reducing the market power of commercial firms.

We say a firm has “market power” when it makes decisions about the price at which it sells its good, not just about the quantity to sell at a given price. Major software manufacturers clearly have enormous market power, measured by the gap between the sales price of packaged software and its cost of production. The cost of copying the software and documentation to disk and packaging it for sale is nugatory, but the package may sell for tens, hundred, or even thousands of US dollars. Economists argue that such market power produces (static) economic inefficiency: the high price means that some consumers do not obtain the good although they are willing to pay more than it costs to produce. From this perspective, the availability of FOSS at its replication cost is a clear efficiency gain over commercial software. This gain may justify government support (e.g., production subsidies or FOSS oriented research grants).

However the cost of additional production—the “marginal” cost of production—does not include the total costs of bringing a good to market. A technology firm will typically have sunk R&D expenditures and fixed costs of marketing. In the absence of market power, the firm would not be in a position to recover such costs. It is typical of technology firms that they incur huge sunk costs to bring goods to market, and under current institutional arrangements,

anticipation of a sales price much greater than marginal production cost is a precondition for their production activity. As a result, market power is often considered a necessary evil, tolerated as the precondition for the production of innovative technology.

Some commercial software faces direct competition from FOSS offerings, which can reduce the market power of the commercial firms. Such competition might increase the incentive for commercial firms to invest in R&D, as they struggle to distinguish their product from the available FOSS offerings. Alternatively, by reducing the anticipated profits of these firms, it may reduce their incentive to invest in R&D. Research on these issues is needed to help determine the effects of FOSS development on software innovation and to guide governments considering active support of FOSS. Unfortunately, research in this area is likely to be tendentious.

As a matter of political economy, we expect commercial firms to finance public criticism of policies that promote or directly subsidize any product that competes with their own. These criticisms will be most persuasive if the subsidized technology is largely imitative, relying parasitically on the technological advances of commercial firms, and if the innovation effort of commercial firms falls in response to competition. So political economists should expect commercial firms to claim FOSS is imitative, despite evidence that many FOSS offerings are technologically innovative. Political economists should also expect to hear claims that FOSS hurts commercial innovation, despite lack of evidence. (Indeed, casual empiricism suggests that FOSS offerings can goad commercial firms to seek technological advances.) Since FOSS policy is inherently political and appears to pose threats to some major commercial manufacturers, political economist should expect that public discussion often will reflect a conflict of

interests rather than the pursuit of understanding.

Government support for FOSS development is generally controversial, but the controversy increases when the development takes place under the GPL. So far, only a small amount of software development underwritten by government has been released under the GPL, but the debate is active. One perspective is that government funding of GPL software creates inefficiencies by disadvantaging the commercial sector, which may find the GPL unacceptable. Another perspective is that government funded research should add to the “knowledge commons” in a way that avoids capture by commercial interests. Dual licensing may offer a partial reconciliation of the two perspectives: government financed software development released under the GPL can also be distributed under licenses more acceptable to commercial firms, in exchange for licensing revenues.

A core policy concern should be that government funded software development produce as much social surplus as possible. Better evidence of the effects on software innovation would help resolve this conflict. However additional complexities lurk in this decision. If government underwritten software development is to be released under the GPL, commercial firms may avoid developing it. Or researchers may withhold revelation of commercially valuable innovations with the intent to commercialize them after completion of their funded research. This may slow innovation when the policy intent is to speed it. Alternatively, some developers may be loathe to work on software that can be hijacked by commercial interests. The possibility of hijacking also influences the revelation incentives in government funded research. These issues, which are likely to remain contentious for many years, would be less important if

software patentability were strictly limited and reduced in duration. Patent policy, however, has been moving in the opposite direction.

Government Use of FOSS

Government use of FOSS is spreading, raising the issue of the proper criteria for government software selection. Some economists have argued that government agencies should simply use total-cost-of-operation, or related criteria, when making software selection (Bessen 2002, Evans 2002). This is often supposed to put FOSS and commercial software on an even playing field. However, the total cost of operation may vary with government policy, especially in the presence of network externalities. In addition, Lessig (2002) correctly notes that government should be interested in the external effects of its software choices.

An increasing number of policy makers appear acutely aware of these issues. Many national governments have adopted policies that are FOSS friendly. Most of these policies permit or encourage the use of FOSS by the government, and the primary motivations appear to be cost saving, compliance with international IP agreements, and promotion of domestic software expertise.

Some Recent Examples

While Europe appears to be drifting toward software patentability along US lines—which may pose substantial threats to FOSS development because it denies the relevance of independent invention—individual European governments have adopted FOSS-friendly policies. In 2003, the British government announced an open source software policy, which attempts to establish a value-for-money standard in the choice of software, and initiated trials of open source software. In June 2002, the German government announced intentions standardize

on FOSS at the federal, state and communal levels. Announcements in 2003 and 2004—primarily a series of GNU/Linux adoption decisions at every level of government—demonstrated that these intentions are being implemented.

Latin America has also signaled receptiveness to a role for FOSS in government. Peru drew particular attention in 2002 when a congressman wrote an open letter to Microsoft defending Peru's "Free Software in Public Administration" bill and incisively critiquing the set of arguments Microsoft had offered against it. In 2002, Venezuela announced that all software developed for the government must be licensed under the GPL. Venezuela linked its adoption of open source to the elimination of software piracy and to a desire to direct government software expenditures to domestic programmers. In 2003 Brazil's president appointed an open source enthusiast to head its National Information Technology Institute, which promptly began to promote government use of GNU/Linux in order to achieve cost savings.

The Middle East has so far seen relatively few governments choose to actively support FOSS development. In 2004, Bahrain chose GNU/Linux for its e-government infrastructure. The government of Israel is promoting open source software in order to cut costs and thereby expand computer use by the public. Israel has also helped develop a Hebrew language version of OpenOffice, enabling the replacement of Windows by GNU/Linux on the desktop.

Many Asian governments have been adopting FOSS friendly policies. China and India have openly expressed support for FOSS, with China emphasizing security concerns and India emphasizing cost concerns. In 2003, the South Korean government announced plans to put open source software on 20%-30% of its PCs. In

2003, the Thai government developed a “people's PC project”, which arranged the mass availability of very inexpensive computers. Cost initially dictated that these be loaded with open source software, and Laser Computer (which sells only GNU/Linux-based PCs) became Thailand's top PC seller. In response, Microsoft cut its prices for a Windows/Office package from nearly \$600 to less than \$40 in Thailand and began development of a localized, reduced-functionality version of Windows XP. The Malaysian government's PC Gemilang Project, which drew a similar price-cutting response from Microsoft in 2004, offers low cost computers running either GNU/Linux or Windows. In 2003, Vietnam announced that it will turn to open source software as a way to reduce its software piracy rates in order to conform to trade agreements it has signed with the U.S. Vietnam's Ministry of Science and Technology wants all state-owned companies and government ministries to switch quickly to open source software, and it is distributing computers loaded with FOSS to schools.

This subsection has highlighted selected national developments, but many municipalities are taking separate action. For example, one of the strongest government actions in support of open source use took place in Australia: as of 2003, the Australian Capital Territory explicitly requires selection of open source over proprietary software “as far as practicable”.

Influences on FOSS Policy

All governments have institutionalized public policies that affect FOSS development. Primary among these are copyright and patent regimes, which affect FOSS development without specifically targeting it. Increasingly governments are adopting more FOSS specific policies. It is intriguing that many of these policies have been FOSS friendly.

Public policy is not made in a vacuum. Social scientists often characterize policy regimes as the outcome of interests competing subject to existing institutional constraints. Benevolence toward others may be one of these interests. Vested commercial interests are always expected to influence policy.

For example, Microsoft Corporation is reportedly lobbying strenuously at local, national, and international levels against laws supportive of open source software. Microsoft lobbied to prevent a World Intellectual Property Organization (WIPO) meeting that was to have considered FOSS software from an intellectual property perspective. The director of international relations for the U.S. Patent and Trademark Office offered a startling critique: “To hold a meeting which has as its purpose to disclaim or waive such [intellectual property] rights seems to us to be contrary to the goals of the WIPO.” Here a specialist in intellectual property rights appears ignorant of the important role that intellectual property plays in FOSS development. It is natural for the social scientist to suspect that such ignorance is willful, and that commercial lobbying rather a vision of the social good is driving this willfulness.

Can FOSS developers compete again well-funded commercial lobbyists who promote policies friendly to well-established proprietary software? This may seem unlikely, but there appears to be remarkable grassroots support for FOSS. The nature of the FOSS community, discussed above, may prove a crucial influence on public policy.

Patent policy may prove a test case. Many FOSS supporters believe that the loose standards for software patentability that have evolved in the U.S. pose a danger to FOSS development, while large commercial software manufacturers tend to favor the U.S. patent regime. Europe's turn of the century

debates over software patentability therefore provide an interesting case study of the influence of competing interests on public policy. The European Union has come under pressure from a variety of groups—including including multinational technology firms as well as the U.S. and the U.K. governments—to allow broad software patentability. The European Patent Office appears eager to accommodate these interests, and has issued broad software patents. However these patents are hard to enforce in the member states, whose laws forbid patents on software or business processes. In response to this situation, patent-favoring interests tried to push through the European Parliament a directive supporting broad software patentability. The directive was passed by the European Parliament in September 2003, but something remarkable happened before passage. Small and medium size technology firms (who believe their businesses are threatened by broad software patents and the large patent portfolios of major technology firms) combined with FOSS supporters (who suspect that broad software patents will be used in efforts to shut down the FOSS community). This coalition successfully pushed for an amendment that strictly limited software patentability. They also successfully added an interoperability exemption, which roughly states that whenever software technology is used solely to facilitate data format conversion, that use will not be considered infringing. (So, for example, a word processor or spreadsheet manufacturer cannot use a file format patent to prevent other applications from reading its files.) These amendments are intriguing in that they do not reflect the interests of the largest players, suggesting that the smaller players stand a chance of being heard in legislative bodies. However the EU Commission and EU Council appear disinclined to bow to the popular will in this matter, and so as of mid-

2004 the final outcome for software patents in Europe remains to be determined.

Conclusion

FOSS development attracts both academic and political attention because of its many intriguing facets. Academics were initially drawn by the innovative creation of economic value outside of traditional market structures. Substantial academic attention has focused on the organizational structures evolved by FOSS developers and on the motivations of individual participants; less has been devoted to the equally important entry of large commercial firms into FOSS development. Early studies uncovered a multitude of policy ramifications of FOSS development, and optimal government policies toward software choice, software underwriting, and software patent policy are yet to be determined. Political aspects of government policy toward FOSS provide additional subject matter for social scientists, and competing interests have already rendered such policies highly controversial. The early 21st century looks to be a turning point, where governments first began to implement policies in an effort to influence FOSS usage and development. The implications for software access and innovation remain to be seen.

Selected References

- Bessen, James (2002). "What Good is Free Software?", in Robert W. Hahn (Editor), *Government Policy toward Open Source Software*, pp. 12–33. Washington, DC: AEI-Brookings Joint Center for Regulatory Studies.
- Cohen, W.M. and D.A. Levinthal. (1989) "Innovation and Learning: Two Faces of R&D", *Economic Journal*, Volume 99, pp. 569–96.
- Evans, David S. (2002). "Politics and Programming: Government Preferences for Promoting Open Source Software", in

- Robert W. Hahn (Editor), *Government Policy Toward Open Source Software*, pp. 34–49. Washington, DC: AEI-Brookings Joint Center for Regulatory Studies.
- Free Software Foundation. (1991) *GNU General Public License*.
- Hahn, Robert W. (2002) “Government Policy Toward Open Source Software: An Overview”, in Robert W. Hahn (Editor), *Government Policy Toward Open Source Software*, pp. 1–11. Washington, DC: AEI-Brookings Joint Center for Regulatory Studies.
- von Hippel, Eric and George von Krogh. (2003) “Open Source Software and the 'Private-Collective' Innovation Model: Issues for Organization Science”, *Organization Science*, Volume 14, Number 2, pp. 209–223.
- Isaac, Alan G. and Walter G. Park. (2004) “On Intellectual Property Rights: Patents vs. Free and Open Development”, in Enrico Colombatto (Editor), *The Elgar Companion to the Economics of Property Rights*. Aldershot, UK: Edward Elgar.
- Lerner, Josh and Jean Tirole. (2001) “The Open Source Movement: Key Research Questions”, *European Economic Review*, Volume 45, pp. 819–26.
- Lerner, Josh and Jean Tirole. (2002) “The Simple Economics of Open Source”, *Journal of Industrial Economics*, Volume 50, Number 2, pp. 197–234.
- Lessig, Lawrence. (2002) “Open Source Baselines: Compared to What?”, in Robert W. Hahn (Editor), *Government Policy toward Open Source Software*, pp. 50–68. Washington, DC: AEI-Brookings Joint Center for Regulatory Studies.
- Office of the e-Envoy. (2002) *Open Source Software: Use within UK Government*. London: Cabinet Office.
- OSI. (Open Source Initiative) (2004) *OSI Approved Licenses*.
www.opensource.org/licenses
- OSI. (Open Source Initiative) (2004) *The Open Source Definition*.
www.opensource.org/docs/definition.php
- Raymond, Eric S. (1998) “The Cathedral and the Bazaar”, *First Monday*, Volume 3, Number 3.
www.firstmonday.dk/issues/issue3_3/raymond
- Shapiro, Carl. (2000) “Navigating the Patent Thicket: Cross Licenses, Patent Pools, and Standard Setting”, *Innovation Policy and the Economy*, Volume 1, pp. 119–50.
- Smith, Bradford L. (2002) “The Future of Software: Enabling the Marketplace to Decide”, in Robert W. Hahn (Editor), *Government Policy toward Open Source Software*, pp. 69–86. Washington, DC: AEI-Brookings Joint Center for Regulatory Studies.
- UNCTAD. (2003) *E-commerce and Development Report 2003*. New York: United Nations.
www.unctad.org/en/docs/ecdr2003_en.pdf
- Villanueva Nuñez, Edgar David. (2003) “Peru Answers MS FUD” (in Spanish).
www.opensource.org/docs/peru_to_ms_spanish.php
- Weber, Steven. (2004) *The Success of Open Source*. Cambridge, MA: Harvard University Press.

Alan Isaac
Department of Economics
American University
Washington DC, USA

Ozone Layer

Jack Reardon

Introduction

Our sun, 150 mio. kilometers away, obtains its energy from nuclear fusion reactions. The energy comes to us as heat and light, necessary for life on earth; but also as harmful radiation, which can potentially destroy all life. Fortunately, the earth's atmosphere has a protective ozone layer that effectively filters harmful radiation from the sun. Without it, life would be impossible. Recent human activity however, has weakened the ozone layer with potentially catastrophic consequences. This has led to successful international cooperation to ban the use of ozone depleting substances. An important lesson learned is that although technology is necessary to raise living standards, the human race must remain ever-vigilant.

Ultraviolet Radiation

All electromagnetic radiation, including light, is composed of tiny photons (from the Greek word 'light') traveling through space at 299,742,458 meters per second, the speed of light—the c in Einstein's famous equation, $E=mc^2$. Electromagnetic radiation differs by wavelength—the distance between two waves—and frequency—the number of times a wave passes a certain point. Wavelength and frequency are inversely related: longer wavelengths have lower frequencies. One of nature's most spectacular sights is a rainbow. Caused by the refraction of sunlight by water droplets, the rainbow displays the spectrum of visible light: red, orange, yellow, green, blue, indigo and violet. The beauty of a rainbow, however, underscores the limitation of the human eye: light, with wavelengths between 400 and 700 nanometers (one nanometer equals one-billionth of a meter) is the only

type of electromagnetic radiation that is visible. Within the spectrum of light (spectrum was coined by Isaac Newton from the Latin 'specter' meaning appearance) red has the longest wavelength and lowest frequency, while violet has the shortest wavelength and highest frequency.

Electromagnetic radiation with wavelengths greater than red, such as infrared, radio waves, and TV waves are invisible; likewise, electromagnetic radiation with wavelengths less than violet, such as ultraviolet (ultra is Latin for beyond), x-rays and gamma rays are also invisible.

A wave's energy is inversely related to its wavelength and directly related to its frequency: thus, electromagnetic radiation with long wavelengths such as radio waves have less energy than electromagnetic radiation with shorter wavelengths such as microwaves. While radio waves can transport beautiful music hundreds of miles, we need the concentrated energy of microwaves to cook our food.

The sun obtains its energy from the fusion of hydrogen atoms into helium, which is lighter than the constituent hydrogen atoms. We know from Einstein's famous equation $E=MC^2$ that mass and energy are interchangeable; thus a small change in mass will generate a huge amount of energy, since it is multiplied by the square of the speed of light. The energy from the nuclear fusion of hydrogen into helium travels to the earth as heat and light, necessary for life on earth, but also as ultraviolet radiation (UV) which is potentially destructive of life.

The destructiveness of UV radiation depends on its wavelength: shorter wavelengths have more energy and are more damaging. Three types of UV radiation have been identified: UV-A with wavelengths between 320 and 420 nm; UV-B with wavelengths between 280 and 315 nm; and UV-C with wavelengths between 100 and 280

nm. UV-A is the least harmful and is used in tanning booths, although excessive exposure can cause cataracts and skin cancer. UV-B can burn exposed skin and can cause melanoma and cataracts. UV-C, the most dangerous, can easily penetrate human skin and alter DNA, causing cancer and death.

What is Ozone?

Oxygen is a molecule—the smallest part of a substance that still retains its chemical composition—of two oxygen atoms (O_2) while ozone is a molecule of three oxygen atoms (O_3). Oxygen and ozone are allotropes (from the Greek ‘other manner’) comprised of the same atoms but with different connecting bonds. Diamonds and graphite, for example, are well-known carbon allotropes.

Oxygen, odorless and colorless, comprises 21 percent of the earth’s atmosphere and 51 percent of the earth’s crust. It easily reacts with most elements to form molecules, such as carbon dioxide (one carbon atom bonded to two oxygen atoms) and water (two hydrogen atoms bonded with one oxygen atom).

Ozone is much rarer than oxygen, comprising one percent of the atmosphere. It is a blueish gas with a sharp, pungent smell. The ancients observed the odor after a lightning strike. Homer, for example, wrote in the *Odyssey*, “Zeus hit the craft with a lightning-bolt and thunder. Round she spun, reeling under the impact, filled with reeking brimstone” (12: 447-449).

In 1840 the Swiss chemist Christian Schoenbein identified and named ozone (O_3) as a separate molecule (from the Greek ‘ozine’ to smell). Ozone’s role in blocking ultraviolet light was discovered in 1879 and two years later the ozone layer was discovered. The ozone layer is not actually a layer but a scattering of concentrated ozone molecules approximately 19-30 km high in the stratosphere, a section of the atmosphere

15km – 50 km above the earth’s surface, as depicted in Diagram 2 (see *Diagram 2: The Earth’s Atmosphere*). Ninety percent of the earth’s ozone is contained in the stratosphere in concentrations between 100 and 300 part per million, while ten percent of the ozone occurs at ground-level between 0.01 and 0.03 parts per million.

In 1924, a Swiss scientist, Gordon Dobson, first measured ozone and established Dobson networks to monitor ozone around the world. Today we use Dobson Units to measure ozone: One Dobson Unit is 0.1mm thick at standard temperature and pressure (zero degrees Celsius at sea level). If the stratospheric ozone was compressed at ground level it would comprise a layer only 3 mm thick.

Ozone can be commercially made by subjecting oxygen to an alternating current electric discharge. Ozone is widely used in industry as an oxidizing agent to clean drinking water, decompose sewage, kill yeast, mold and bacteria on fruits and vegetables.

Formation of the Earth’s Ozone Layer

Our solar system was formed 4.5 billion years ago as a result of a supernova explosion, which occurs when a star’s fuel is completely consumed forcing it to collapse inwards, resulting in huge release of energy and materials from its ‘ash heap.’ After millions of years, the earth cooled sufficiently, forming reservoirs of water, which effectively shielded solar radiation and allowed primitive algae to develop. The algae used sunlight to decompose carbon dioxide (CO_2) into carbon for energy while releasing oxygen as a waste-product. This gradually increased the atmosphere’s oxygen content while simultaneously decreasing the level of carbon dioxide.

Approximately 600 million years ago, the stratospheric ozone layer was formed, shielding UV radiation and allowing more

complex multi-cellular plants to develop. Central to the formation of the ozone layer was the interaction between highly energetic UV radiation and stratospheric oxygen. UV radiation, traveling at the speed of light, decomposes diatomic oxygen (O_2) into two separate oxygen atoms. An atom is mostly empty space. Protons and neutrons with a positive and neutral charge respectively, comprise the nucleus, while electrons, with a negative charge inhabit shells surrounding the nucleus. The electron shell furthest from the nucleus (where the electromagnetic repulsive force between protons and neutrons is weakest) determines the atom's bonding potential. If the outermost shell is filled with electrons and has no more available space, it is inert and cannot bond. Examples include the inert gases: helium, argon, neon, and radon. If the outermost shell of an atom has empty space, it will bond with other atoms or molecules in order to create a more stable molecule with a lower energy level.

Since a single oxygen atom has empty space in its outermost electron shell it will readily bond with an oxygen atom (O_2) to form ozone (O_3).

Ultraviolet radiation also destroys ozone, by separating ozone into oxygen (O_2) and single oxygen atoms. The heat released from the decomposition of ozone warms the temperature of the stratosphere causing it to increase with altitude; unlike the troposphere, where temperature declines one degree Centigrade for every 100 meters ascended. Because of the temperature inversion, stratospheric gases (from the Greek 'stratos' meaning layered) do not mix and air flows horizontally.

While stratospheric ozone is beneficial, ozone at ground level is hazardous. It is formed when nitrogen oxide (NO_x) from automobile emissions and coal-fired utilities interact with heat, sunlight and other air pollutants. Nitrogen oxide results from the

oxidization – the mixing with oxygen during combustion. Once released into the atmosphere, nitrogen oxide combines with a single oxygen atom to form nitrogen dioxide (NO_2). On hot sunny days, ultraviolet radiation decomposes NO_2 back into NO_x releasing a highly reactive single oxygen atom, which readily latches onto diatomic oxygen (O_2) to form ozone. Ground-level ozone is the principal component of smog and can trigger asthma attacks, increase the risks of respiratory infections and permanently damage lung capacity.

The existence of the ozone layer underscores the delicate conditions necessary for life on earth. It is indeed the "rarer gases [that] tend to be major participants in the business of life" (Lovelock 2000, 73). While too little ozone is catastrophic, too much ozone can limit UV and lead to a Vitamin D deficiency, which relies on exposure of the skin to UV radiation (Lovelock 2000, 71). In the stratosphere, before human intervention, ozone molecules were created and destroyed at a constant rate (Chapman 1930). But human interaction has adversely affected the ozone layer jeopardizing its resiliency.

Human Destruction of the Ozone

One of the more remarkable inventions in human history is the refrigerator (from the Latin 'frigus' meaning frost). It works by circulating a material within an enclosed area to absorb heat from the surrounding ambience and then ejecting outside in gaseous form.

Ammonia (NH_3), methyl chloride (CH_3Cl) and sulphur dioxide (SO_2) were early refrigerants, but after a series of fatal accidents involving leakages of these toxic gases, a collaborative effort was undertaken by DuPont and General Motors to develop a safer alternative. In 1928, Thomas Midgely, a research chemist with General Motors and the inventor of leaded gasoline, synthesized chlorofluorocarbons (CFCs) as a viable

substitute. CFCs were named for their molecular configuration of carbon, fluorine and chlorine. Hailed as a wonder gas because it is non-toxic, non-corrosive, nonflammable and inert, CFCs were given the tradename Freon by Dupont and became widely used as coolants in refrigerators and air conditioning systems. CFCs were later used in aerosol sprays, fire extinguishers, Styrofoam, insulation, foaming agents and as solvents in the electronics industry. In fact, by the late 1980s more than 250 product categories were made from CFCs and related substances; some frivolous, but others essential to industrial society (Anderson and Sarma 2002, 189-190).

The desirable properties that made CFCs a wonder gas, however, also prevented their decomposition in the atmosphere. In a pathbreaking paper, Professors Rowland and Molina (1974) demonstrated that ultraviolet radiation can decompose CFCs in a chemical reaction, in turn destroying stratospheric ozone.

Ozone destruction occurs in three stages, with chlorine playing the key role as catalyst, i.e, it is instrumental to the reaction without itself being destroyed. Chlorine is a member of the halogen family (from the Greek 'salt-producing'). Think of common table salt, also known as sodium chloride. The halogens: fluorine, chlorine, bromine, iodine and astatine, share common chemical properties and are thus grouped together in the Periodic Table of the Elements. One common property is their reactivity. Since all halogens have unfilled space in their outermost electron shell, they will readily bond. Smaller atoms tend to be more vigorous in forming bonds; thus fluorine and chlorine, the two smallest halogens are the most reactive. Once halogens bond however, only a great amount of energy, such as ultraviolet radiation can decompose the molecule.

In the first stage of stratospheric ozone destruction, highly energetic UV radiation decomposes the strongly bonded CFCs, releasing single chlorine atoms. Next, a single chlorine atom bonds with ozone to form chlorine oxide (ClO) and oxygen (O_2). Finally, chlorine oxide interacts with a single oxygen atom to form one chlorine atom and oxygen (O_2). Chlorine destroys ozone, while emerging intact to destroy more ozone molecules. One chlorine atom in the stratosphere has a life expectancy between 4 and ten years can destroy 100,000 ozone molecules, before bonding with hydrogen to form water-soluble hydrogen chloride, allowing it to dissipate into the troposphere.

Midgely thought he had produced a wonder gas without deleterious economic consequences. It is thus ironic that he has posthumously been called the "organism with the most significant impact on earth's history" (McNeil 2000, 111). Midgely's inventions underscore the dual nature of technology which can provide solutions to pressing problems thereby improving living standards, yet also engender unforeseen deleterious consequences.

Although CFCs are the most familiar ozone depleting substance, others, which use either chlorine or bromine, include carbon tetrachloride and methyl bromide, used in fire extinguishers and pesticides and methyl chloroform.

CFCs also contribute to global warming. In addition to naturally occurring greenhouse gases that naturally occur such as carbon dioxide, methane, nitrous oxide and tropospheric ozone, CFCs, all artificially made, also contribute to global warming, principally CFC₁₁, CFC₁₂ and CFC₁₃. A word here on the numbering system of chlorofluorocarbons. An easy way to understand the meaning of the subscript is the rule of 90, whereby 90 is added to the subscript (Elkins 2002). For example, with

CFC₁₁, add 90 to obtain 101. The hundreds digit represents the number of carbon atoms; the tens digit represents the number of hydrogen atoms and the ones digit represents the number of fluorine atoms. So with CFC₁₁ we have 1 carbon atom, no hydrogen atoms and one fluorine atom. The number of chlorine atoms is obtained from the equation, $Cl = 2(C + 1) - H - F$. So with CFC₁₁, we have $2(1 + 1) - 0 + 1 = 3$ chlorine atoms.

Global warming slows the ability of the ozone layer to repair itself, because the heat trapped in the lower atmosphere, radiates to the upper atmosphere where it forms ice crystals, which as mentioned earlier, acts as a catalyst in ozone destruction.

In addition to CFCs, nitrogen oxide, an emission from the supersonic transport (SST) was identified as destroying the ozone layer and was a major reason for the cessation of the SST program in many nations. The detonation of a nuclear bomb releases nitrous oxides into the atmosphere and can initiate the ozone depleting process (Schnell 1982). In addition to the catastrophic destruction of the bomb itself, depletion of the ozone layer is another reason to prevent the usage of nuclear weapons.

Global Solutions to Ozone Depletion

Since the development of agriculture 10,000 years ago, a distinguishing characteristic of human societies has been the outstripping of resources and increased pollution (Pointing 1991:383). Before the Industrial Revolution environmental problems were localized requiring local solutions. Economic growth and new technology spawned by the Industrial Revolution, however, has given rise to many global environmental problems. Two of them – global warming and depletion of the ozone layer, have been called “the most serious pollution threats the world has ever faced” (Pointing 1991, 383).

These two global problems are also linked in origin. Ancient plants extracted carbon dioxide to obtain energy for growth, while exhaling oxygen as waste, contributing to the formation of the ozone layer. Over millions of years, the algae and plants became buried under layers of sediments. They were heated and pressurized into coal, oil and natural gas – today’s fossil fuels, the combustion of which is the preponderant factor in global warming.

Sceptics claim that the earth has experienced natural climate changes and periodic natural ozone changes, citing Europe’s mini ice age (1200 AD – 1800s) and significant ozone reductions from the eruption of Mt. Pinatubo in the Phillippines in 1991, that emitted chlorine directly into the stratosphere. Nevertheless, incontrovertible evidence exists that human activity has contributed to both global warming and ozone depletion, which if unchecked, will cause catastrophic and irreparable damage.

After the Rowland and Marina seminal report, the Governing Council of the UN Environmental Program officially discussed ozone depletion in 1976. One year later, the United States, the world’s largest emitter of CFCs unilaterally banned non-essential aerosol products used mainly for cosmetic purposes, followed swiftly by Canada, Sweden and Norway. This led to the Vienna Convention for the Protection of the Ozone in 1985 to encourage intergovernmental cooperation on research and systematic monitoring of the ozone layer.

In 1984 a huge hole in the ozone layer was discovered over Antarctica directly linked to CFCs. Ozone depletion is particularly severe over Antarctica during the winter months because extreme cold temperatures cause the formation of polar stratospheric clouds. The clouds, composed of nitric acid and ice provide an ideal catalytic surface for ozone depletion, which is not as severe over the

Arctic given the relatively warmer temperatures.

With ozone depleting substances expected to triple, a fifty percent reduction in the ozone layer was forecast by 2035, with each one percent reduction leading to a two percent increase in the incidence of cataracts, blindness (particularly severe among animals since they cannot wear protective sunglasses) and skin cancer. Phytoplankton, the unicellular organisms at the base of the aquatic food chain, inhabiting the ocean's upper layer would be most susceptible. Ozone depletion would also increase damage to other marine life and stunt photosynthesis.

Given the urgency of the problem, nations around the world recognized the need for stronger action. In 1987, 24 nations and the European Community signed the Montreal Protocol, which has since been signed by 184 nations. The Protocol initially mandated a fifty per cent reduction in CFCs, and was renegotiated in London (1990), Copenhagen (1992), Vienna (1995), Montreal (1997) and Beijing (1999), as additional scientific evidence proved that ozone depletion was occurring faster than expected.

The essence of the Protocol and its extensions is to ban substances containing either chlorine or bromine. The Protocol grants minor exceptions for the continued use of CFCs including feedstocks in which ODS are consumed; process agents in which emissions are tightly controlled and medical applications such as cleaning pace makers and artificial limbs. The Protocol does not mandate how each country will achieve the mandate; thus, a wide range of methods have been used including taxes, subsidies, quotas, regulations and market trading permits.

Assuming full enforcement of all current provisions, the ozone layer will return to normal around 2050 and the Antarctica ozone hole will disappear. Ozone depleting compounds peaked in 1994, and since 1986,

the total production and consumption of CFCs has decreased by more than 85 percent. The number of skin cancer deaths, however, will continue to increase before leveling off. Nevertheless, without the Protocol ozone depletion would be ten times greater today, with millions more cases of skin cancer and death and possibly irrevocable damage to agriculture and ecosystems (Anderson and Sarma 2002, 346).

The global response to the ozone layer problem represents "an extraordinary response to an extraordinary problem" (McNeil 2000:113) as well as "unprecedented cooperation" (Anderson and Sarma 2002:345). Several factors account for the success of the Protocol. The first is the recognition that developed nations have a special responsibility to curtail emissions and assist developing nations. The Montreal Protocol established two different time frames for developed and developing nations. Developed countries were given a ten-year delay as long as their use of CFCs did not increase significantly. CFCs, for example, were phased out in developed countries in 1995 and 2010 in developing countries; halons in 1993 and 2010 respectively; carbon tetrachloride in 1995 and 2010; HCFCs in 2020 and 2040 and methyl bromide in 2005 and 2015.

The second is the assistance given to developing nations as a Multilateral Fund was created in 1992 to provide funds for technological transfer and assistance in adopting new technologies. It is the only such fund to be implemented in a multilateral treaty.

The third is the development of substitutes for CFCs which tempered the reluctance of industry to cooperate. Principal substitutes include hydrochlorofluorocarbons (HCFCs) and hydrofluorocarbons (HFCs). The former include hydrogen atoms which react with other atmosphere gases, thus reducing their

lifespans to about 13 years on average compared to CFCs with lifespans between 10 and 100 years. The Copenhagen Agreement calls for HCFs to be eliminated by 2030. HFCs, on the other hand, do not contain chlorine and thus do not destroy the ozone and have been adopted in auto and office air conditioners.

The fourth is the widespread acceptance of the evidence and the urgency of the problem. This led to widespread cooperation between science, government, NGOs and industry, although initial industry recalcitrance delayed earlier agreement. During the first ten years after the Rowland and Marina report, for example, industry argued for more conclusive research, a position not unlike American corporations today regarding global warming.

The final factor is the perseverance of governmental agencies and individuals in maintaining the momentum for the passage of some type of international agreement, despite early industry resistance. It was passage of the Montreal Protocol that expedited the development of viable substitutes such that “regulations forced the development of technology” (Anderson & Sarma 2002, 362).

Can the success of the Protocol be replicated to solve global warming? Yes, if the false and misleading dichotomy between command and market approaches is jettisoned. The command approach involves a government-issued edict that bans a pollutant without giving discretion to market participants to implement the assumed necessary technology. Market-based approaches set overall limits to pollutant and allows market participants the discretion as to how best to reach the overall goal.

But the dichotomy between command and market is misleading, since both approaches are “man-made rules that govern behavior” (Swaney 1992:627). Although the command approach has recently fallen out of vogue in favor of a market approach “(i)t is a myth to

cast taxes and regulations as polar opposites ... The full truth is that the approaches are best seen as compliments not rivals ... Making the industrial economy operate efficiently within environmental limits will require synthesizing the two approaches, using the strengths of each to compensate for the weaknesses of the other” (Roodman 1997:26).

A command approach was justified as a solution for the ozone problem, given the widespread acceptance of the urgency of the problem, the relative ease in producing substitutes, and the acknowledgment that not all costs could be internalized. Whereas, a major obstacle in securing a universal international agreement on global warming is the recalcitrance of the coal industry, especially in the United States, which accounts for one-fourth of global production. Although a command approach was implemented in the Kyoto Treaty mandating a reduction of carbon dioxide emissions, it was rejected by the United States in favor of a market-based system. This recalcitrance can best be understood as lack of alternatives to coal. Specifically, there is no technology to curtail the emissions of carbon dioxide – a preponderant emission from the combustion of coal, other than curtailing its use. Thus, even if the use of coal were “regulated even modestly, [it] would change the future of coal use entirely” (Freeze 2003:182).

Conclusion

There will be more ozone problems associated with new technology. Changes in technology make progress possible, but also render current rules of the market inadequate; thus they create the need for institutional change (Swaney 1992:626). And there will be future Thomas Midgelys: people who honestly think they are developing a wonder product only to learn years later of unforeseen environmental consequences.

A lesson from the ozone problem is that humanity “should be vigilant forever to ensure that ozone depleting substances are banished from the world” (Anderson and Sarma 2002:367). Global institutions should foster a more equitable sharing of resources and a global network to vigilantly monitor the effects of technology on the ozone layer.

Selected References

- Anderson, Steven and K. Madhava Sarma. (2002) *Protecting the Ozone Layer*. London: Earthscan Publications.
- Benedick, R.E. (1998) *Ozone Diplomacy*. Cambridge, Mass.: Harvard Univ. Press.
- Cagin, S. and P. Dray. (1993) *Between Earth and Sky: How CFCs Changed our World and Threatened the Ozone Layer*. New York: Penguin.
- Chapman, Sydney. (1930) “A Theory of Upper-Atmospheric Ozone.” *Memoirs of the Royal Meteorological Society*, Volume 26, pp. 103-25.
- Elkins, James. (2002) “Chlorofluorocarbons.” Climate Monitoring and Diagnostic Laboratory. www.cmdl.noaa.gov
- Farman, J.C.; B.G. Gardiner and J.D. Shanklin. (1985) “Large Losses of Total Ozone in Antarctica Reveal Seasoned ClO_x/NO_x Interactions”, *Nature*, Number 315, pp. 207-10.
- Freese, Barbara. (2003) *Coal—A Human History*. Cambridge, Mass.: Perseus Publ.
- Lovelock, James. (2000) *Gaia: A New Look At Life on Earth*. Oxford, UK: Oxford University Press.
- Matsumura, Y. and H.N. Ananthaswamy, (1953) “Toxic Effects of UV Radiation on the Skin.” *Applied Toxicology Pharmacol*, pp. 298-308.
- McNeil, J.R. (2000) *An Environmental History of the Twentieth-Century World*. New York: W.W. Norton.
- Midgley, Thomas and A. Henne. (1930) “Organic Fluoride as Refrigerants.”

Industrial and Engineering Chemistry, Volume 22, pp. 542-47.

- National Academy of Sciences. (1979) *Protection Against Depletion of Stratospheric Ozone by Chlorofluorocarbons*. Washington DC: NAS.
- Pointing, Clive. (1991) *A Green History of the Earth*. New York: Penguin.
- Roodman, David. (1997) *Getting the Signals Right: Tax Reform to Protect the Environment and the Economy*. World Watch Paper 134. Washington, DC.
- Schell, Jonathan. (1982) *The Fate of the Earth*. New York: Knopf.
- Solomon, S.; R.R. Gareig; F.S. Rowland and D.J. Weabbes. (1986) “On the Depletion of Antarctica Ozone”, *Nature*, Number 321, pp. 755-58.
- Swaney, James. (1992) “Market Versus Command and Control Environmental Policies.” *Journal of Economic Issues*, Volume 26, pp. 623-33.

Websites

- Visual Tour of Ozone Hole. www.atm.ch.cam.ac.uk/tour
- Ozone Helpful Questions. www.theozonehole.com.
- NASA Site for Ozone Education. www.nasa.gov/About/Education/Ozone/OzoneLayer
- Environmental Protection Agency. Environmental Indicators: Ozone Depletion. epa.gov/cgi-bin/epaprintonly.cgi.
- Montreal Protocol. www.unep.org/ozone/pdfs/Montreal-Protocol

Jack Reardon
School of Business
Hamline University
St. Paul,
Minnesota, USA
jreardon02@gw.hamline.edu

Patents and Copyrights

Wilfred Dolfsma

Introduction

Patents and copyrights are the two most important examples of legal means to prevent the unrestricted and unconditional imitation of fruits of the intellect. Other Intellectual Property Rights (IPRs) include trademark law, plant patents and design patents. Secrecy is an important additional means of protecting knowledge from being used by others than the person or organization that has developed it. Secrecy, however, needs to be enforced by other bodies of law such as labor or contract laws, and is thus not part of IPR.

IPRs differ substantially in the kind and degree of protection that they offer, reflecting the different rationales for their existence. Duration, scope and thresholds for protection vary from IPR to IPR, vary over time, and vary between countries. IPR is statutory law, enacted by the legislator, and thus is national law. In general it can be said that the importance of IPRs for the global economy has increased over time. This is reflected in the increase in the number of applications for patents, but also in the inclusion international agreements such as WIPO and TRIPS. Some measure of harmonization has thus resulted in recent years.

What are Patents?

The first patent was probably extended in the 14th Century, while the first 'patent law' dates from 1474 by the Republic of Venice (Machlup 1958). Thus, the patentee was given the exclusive right to produce and sell within the boundaries of its jurisdiction for a limited number of years. Patents offer the most powerful protection from imitation as they protect the newly developed *idea* itself. In contrast to patents, copyright rather protects the particular way in which an idea is

expressed – moderations of the idea that are sufficiently different may be developed by third parties without further ado.

A person or an organization must apply for a patent. Applications are evaluated by patent offices using several criteria. One criterion is that the idea must represent a new, non-obvious development when compared with 'prior art'. Prior art are the applications made and patents granted in the same field in the past. An application needs to offer an inventive step; incremental innovations do not qualify. The application needs to offer the possibility for industrial application. Finally, a physical component must be involved.

All of these criteria, which are used across most developed countries, are imprecise. The extent to which they are further clarified differs over time and across space. This is apparent in the discussions on the issue of whether software and business models warrant protection from patent law. US patent law now offers protection for the both of them, on the argument that both in many cases rely critically and cannot be separated from a physical component. The distinction between hardware and software is, for instance, in many cases difficult to make. This is an issue that relates to the question of the breadth or scope allowed for patent applications. What may be included under a patent; how specific should the use of new knowledge be before a patent can be granted?

There are costs involved in applying for a patent, there are yearly costs for renewal of the application, there are costs to search for possible infringement, and there are costs for associated with the legal action needed to redress infringement. All of these costs need to be born by the applicant. Filing and maintaining a patent is costly, and subject to important procedural difference across jurisdictions that can have profound effects on organizations' position (OECD 1997). Unlike many other countries, the US has a

‘first to invent’ rule; if one is able to show one was first to invent, one can claim priority and be granted the patent even when another party filed for or was granted the patent earlier. The first to invent rule is administratively much more burdensome than the ‘first to file’ rule. Even though in principle it will benefit the actual creative party more, it invites more legal disputes and promotes strategic behavior between firms. If and when proof of development is properly registered a firm would claim to have right on the patent after another firm has filed for it, having committed many resources in the process.

Balancing public and private interests, most patent laws limit the extent to which an exclusive license is given to the right holder. Limiting the duration of patents is one example of this. For patents, an important instance where public interests might overrule private ones in particular cases is in the form of a ‘compulsory license’. The rights of patent holders can be curtailed when society as a whole would benefit as a consequence – in such cases, the patent holder is forced to license the knowledge at reduced or no charge. The production of anti-HIV medication for people in Sub-Saharan Africa who suffer from AIDS is an example. As one can imagine, there is bound to be discussion about when to deem compulsory licenses justified.

What Are Copyrights?

Copyrights apply to ‘original works of authorship’, literary and artistic works in fixed form. Examples include writings, music, drawings, dances, computer programs, movies and now also data bases. Ideas, concepts, principles, (‘brute’) facts and (tacit) knowledge are not included. The first time copyrights were formally enacted was in the English Statute of Anne (1709), based on informal previous practice.

Copyright law protects the *expression* of an idea; one merely has to be able to prove anteriority by proving to have published the work first. Copyrights actually refer to a number of different though related rights to such works. Copyrights are exclusive right to use or authorize others to use the work on agreed terms. It includes reproduction in various forms, recordings of it, its broadcasting, its translation into other languages, or—more controversially—its adaptation, such as a novel into a screenplay. In the latter case, one departs from the idea that copyrights protect the particular expression of an idea. So-called ‘neighboring rights’ have been added, preventing the public performance of a (musical) work without consent from and particularly payment of royalty to the author.

In contrast to patents, one does not have to apply in order for one’s creative work to be protected under copyright law. In some countries such as France one does have to send a copy of the work to a central location for storage. Copyrights can be considered to provide weaker protection than patents do. On the other hand, copyrights’ duration is longer, generally lasting for the life of the author plus 70 years. Patents generally last 20 years. There is an overall increase in the duration copyrights provide protection. The Statute of Anne provided protection for 14 years (once renewable). Since 1993 in Europe, and 1998 in the USA, the term mentioned above holds.

There are a number of limitations as to what may be protected under copyright law; the kinds of limitations and/or the way these are enacted differs across countries. The most important among the limitations is known as ‘fair use’ (‘fair dealing’ in the UK). For private use copies may be made without liability for infringement and payment of royalty. For, for instance, journalistic or

educational purposes, parts of a work may also be copied without infringing copyrights.

The duration and scope of patents, but of copyrights especially has increased over particularly the most recent years to a degree that once objects *per se* ruled out for protection, such as collections of ‘brute facts’ (databases) are now included (Maurer et al. 2001). This increases the costs for parties involved who would use the protected knowledge to be creative themselves. Innovation may well be hampered as a result (Baumol 2002).

Why IPRs are (Not) Needed

Discussion of the need for society of IPRs has waxed and waned. Notwithstanding such discussions, the scope and duration of IPRs has increased steadily over time. Intellectual objects are non-exclusive: consumption or use by non-payers cannot be excluded. In addition, intellectual objects are partly non-rivalrous as well: they are not consumed by their use. This makes intellectual objects (quasi-) public goods, giving governments a reason to influence relevant processes in society. As costs of imitating or communicating intellectual objects tends to be low, there is a tendency for these to be under-produced (see, e.g., Nelson 1959, Romer 2002). IPRs would provide a way to compensate creative individuals that is saving on transactions costs.

Rationales for IPRs fall into four, partly related categories (Hettinger 1989). The extent to which rationales are stressed in law differs between countries, reflected in the authority that administers them. In the UK and the US, the incentive for creative individuals or organizations that IPRs offer is emphasized. Without IPRs one would be less inclined or not inclined at all to develop and diffuse new knowledge. The prospect of a period of time in which one is able to commercially exploit the innovation will, in

this view, offer an incentive to create and diffuse new knowledge. This rationale is founded in John Locke’s argument for property rights in general. In his view, a person’s gains that with which she ‘mixes her labor’, provided that ‘enough and as good [is] left in common for others’.

Relatedly, IPRs are said to be necessary for firms to entice them to invest in facilities for the production of goods based on the intellectual object protected under IPR. Without it, firms would face more than the usual business risk and refrain from the production of goods that would, presumably, benefit society as a whole. In the UK and the US, these are the rationales emphasized, and this is reflected in the fact that the Commerce Department administers such rights.

The two other rationales are not related to such utilitarian considerations and are specifically emphasized in the legal systems of continental Europe (and those based on or influenced by them). The first is one of desert. If someone has produced an intellectual object, he deserves some kind and measure of reward. The final rationale is personal/moral one. In creating an intellectual object, someone expresses one’s personality. The object is part of the self, so to speak. A result of this is that copyrights in a European context include so-called ‘moral’ rights. These are inalienable, non-transferable. Even when a piece protected under copyright law is sold, the new owner may not alter it without consent of the author.

Over time, the first and second rationales have become increasingly dominant in the discussions. Philosopher John Locke’s argument for a natural property right in what one makes has a strong intuitive appeal. In reality, however, a government creates and polices IPRs; they are a socially created privilege. Intellectual objects, in addition, differ from physical ones. In their creation, for instance, one draws on work done (by

others) in the past; creation is not *de novo*. When use of existing work is restricted, society may be hurt. As intellectual objects are public goods, granting a (temporary) monopoly on their commercial exploitation may not leave 'enough and as good', particularly in the case of patents as they protect the idea developed. Independent inventors are hurt as they may be prohibited from using something they have developed themselves but another party developed or was granted a patent for earlier. It is further argued that intellectual objects are more often than physical ones the result of cooperation. These issues make the rationale for having IPRs weaker.

How Important are IPRs?

Baumol (2002) has estimated that twenty percent of the benefits associated with an invention are appropriated by the parties directly or indirectly involved with the invention. Only partly will the appropriation of benefits be due to IPRs. Economically, the significance of IPRs is difficult to establish. Indirect measures will have to be relied on, each having their particular advantages and disadvantages. As noted, the number of patents granted grows exponentially. When works still needed to be registered in the USA in order to be granted protection under copyright law, a similar growth was visible (e.g., Andersen et al. 2000).

The actual advantage for firms of being able to exploit an intellectual object commercially on an exclusive basis is disputed empirically. Levin et al. (1987) present empirical data that indicate that patents are not considered the most important means to protect one's position. This differs, of course, across industries, with firms in industries such as the pharmaceutical industry, where inventing-around is difficult, indicating that patents are important. These findings have been replicated after Levin also

for countries other than the USA (Arundel 2001).

Welfare economic analyses show that the use of patent can be beneficial to society as a whole if and when their breadth is limited, even though their duration may be longer than it is now (see articles reproduced in Towse & Holzhauser 2002). This increases the possibilities for inventing-around a patent that has been granted. Such possibilities are as much dependent on the nature of the technology or patent law, as they are on the strategy of the firms (Grandstrand 1999). Partly, the tremendous growth in the number of patents granted (see OECD 2001) can be attributed to the increasingly strategic nature of patent applications.

What Effects Do IPRs Have?

Eminent 20th Century economist Joan Robinson has said "since it is rooted in a contradiction, there can be no such thing as an ideally beneficial patent system, as it is bound to produce negative results in particular instances". It is clear that the patent system provides incentives for people and organizations to develop and provide intellectual objects to a market. The relevant economic question from a policy perspective is whether a different system would be able to do so more effectively. In other words, what are the opportunity costs of the present system of IPRs? About patents, Machlup (1958:28) has famously remarked that if we did not have any such system, "it would be irresponsible, on the basis of our present knowledge of its consequences, to recommend instituting one. But since we have had a patent system for a long time, it would be irresponsible, on the basis of our present knowledge, to recommend abolishing it." This conclusion is still endorsed. The reason is that patent law, but IPR in general, is to serve several goals. It is to promote creativity, but does so by promising the right

to exclusive exploitation. Exclusive exploitation is not similar to a monopoly, however, for one because the product does not define the relevant market. Several products usually compete on a single market. In return for this exclusive right, the patent holder is—as the TRIPS agreement has stated it—to “disclose the invention in a manner sufficiently clear and complete for the invention to be carried out by a person skilled in the art and may require the applicant to indicate the best mode for carrying out the invention known to the inventor”. Disclosure of knowledge may be a basis for others to further develop knowledge in the field.

There are several assumptions underlying the general idea in favor of IPRs that it stimulates creativity. One is that creative individuals are (primarily) motivated by financial incentives. This may not always be true. For instance, the amounts copyright holders tend to receive in the mean will hardly be an incentive to continue their creative work (Towse 1999). Still, the supply of some copyright protected works does seem to respond to the market price (Hui & Png 2002). This relates to another underlying assumption, which is that there is a creative individual and not a group or even an organization that does the creation (Fisher 2001). The creative *labor* is undertaken by an individual (Menell 2000), who would therefore have a *natural* right in the object. The idea of the lonely and somewhat goofy inventor is a remarkably persistent one. Other ways to provide incentives that, theoretically, work equally well do exist (Shavell & Ypersele 2001). At the same time, in circumstances where multiple individuals have contributed, it would be cumbersome if each of them could prevent the exploitation of the work. Furthermore, the way in which IPRs are implemented (e.g., OECD 1997), or administered (e.g., Kretschmer 2002) may be cumbersome and introduce perverse effects.

IPRs in a Global Age

The need for harmonization of intellectual property rights at a global level has been felt for a long time. A number of conventions have sought to do so. The Berne (1886; amended and revised several times since), Rome (1961) and Phonogram (1971) conventions, to mention the most important ones, have brought a measure of harmonization for copyrights. Under the Uruguay round of the General Agreement on Tariffs and Trade (GATT), started in 1986 and concluded in 1994, a treaty on Trade-Related Aspects of Intellectual Property Rights (TRIPS) has also been adopted which brought further harmonization. Indeed, as this more or less ad hoc organization transformed into a standing World Trade Organization (WTO) in 1995, countries could no longer decide to enter the global arena selectively. Instead, they now have to accept the treaties that the WTO supports, including TRIPS, *in toto*. The World Intellectual Property Organization (WIPO, under the auspices of the United Nations) also ‘seeks to promote the use and protection of works of intellectual objects’. Net exporters of products embodying intellectual property tend to seek to promote the use and promotion of IPRs.

In the past, countries that are now among the wealthy countries of this world (the USA, Japan, the Netherlands) have not signed up on international treaties that would force them to tighten their IPR law if it hurt their interests. Firms or organizations in those countries could and did take the knowledge developed by others—made public, e.g. in order to be granted a patent—to develop products themselves. Developing countries are now deprived of this possibility to ‘free ride’ on knowledge developed by others.

Selected References

- Andersen, A.; J. Howells; R. Hull; I. Miles and J. Roberts. (2000) (Editors) *Knowledge and Innovation in the New Service Economy*. Cheltenham: Edward Elgar.
- Arundel, A. (2001) "The Relative Effectiveness of Patents and Secrecy for Appropriation", *Research Policy*, 30, 611-24.
- Baumol, W.J. (2002) *The Free-Market Innovation Machine: Analyzing the Growth Miracle of Capitalism*. Princeton, NJ: Princeton UP.
- Fisher, W. (2001) "Theories of Intellectual Property", in Stephen Munzer (Editor), *New Essays in the Legal and Political Theory of Property*. Cambridge, UK: Cambridge UP.
- Granstrand, O. (1999) *The Economics and Management of Intellectual Property*. Cheltenham: Edward Elgar.
- Hettinger, E.C. (1989) "Justifying Intellectual Property", *Philosophy and Public Affairs*, 18, 1, 31-52.
- Hui, K.-L. and I.P.L. Png. (2002) "On the Supply of Creative Work: Evidence from the Movies", *American Economic Review*, 92, 2, 217-20.
- Kretschmer, M. (2002) "The Failure of Property Rules in Collective Administration", *European Intellectual Property Review*, 24, 3, 126-37.
- Levin, R.; A. Klevorick; R. Nelson and S. Winter. (1987) "Appropriating the Returns from Industrial Research and Development", *Brookings Papers on Economic Activity*, 3.
- F. Machlup. (1958) *An Economic Review of the Patent System*. Washington DC: US Government Printing Office.
- Maurer, S.M.; P.B. Hugenholtz and H.J. Onsrud. (2001) "Europe's Database Experiment", *Science*, 294, 789-90.
- Menell, P.S. (2000) "Intellectual Property: General Theories", in: B. Bouckaert and G. de Geest (Editors), *Encyclopedia of Law and Economics*. Cheltenham: Edward Elgar, 129-88.
- Nelson, R.R. (1959) "The Simple Economics of Basic Scientific Research", *Journal of Political Economy*, 57, 297-306.
- Organization for Economic Cooperation and Development. (1997) *Patents and Innovation in the International Context*. Paris: OECD.
- Organization for Economic Cooperation and Development. (2001) *Science, Technology and Industry Outlook: Drivers of Growth: Information Technology, Innovation and Entrepreneurship*. Paris: OECD.
- Romer, R. (2002) "When Should We Use Intellectual Property Rights?", *American Economic Review*, 92, 2, 213-6.
- Shavell, S. and T.V. Ypersele. (2001) "Rewards versus Intellectual Property Rights", *Journal of Law and Economics*, 44, 2, 525-47.
- Towse, R. (1999) "Copyright and Economic Incentives: An Application to Performers' Rights in the Music Industry", *Kyklos*, 52, 3, 369-90.
- Towse, R. and R. Holzhauser. (2002) (Editors) *The Economics of Intellectual Property*. 4 Volumes. Cheltenham: Edward Elgar.

Wilfred Dolfsma

Dept of Innovation Management & Strategy
University of Groningen, Groningen

The Netherlands

W.A.Dolfsma@rug.nl

Pensions, Superannuation and Population

Christian E. Weller

Introduction

In many industrialized and industrializing countries, life expectancies are rising and birth rates are falling—leading many observers to conclude that existing retirement systems, which rely on workers to pay directly for retirees, are doomed to fail. The fact that a shrinking number of taxpayers is expected to support a growing number of beneficiaries is often deemed unsustainable in the long-run.

As current retirement systems are projected to encounter severe difficulties in financing promised retirement benefits policymakers are often focusing on two possible paths to improve the outlook of their countries' retirement systems. For one, promised benefits are reduced to help improve pension finances, e.g. by raising the retirement age. Additionally, policymakers are shifting the financial burden of saving for retirement onto individuals, while reducing future liabilities for governments and employers. For instance, many countries are proposing to create personal accounts that are intended to replace part or all of public pensions.

From a policy perspective, problems arise due to two divergent trends. On the one hand, the need for secure retirement benefits is growing with aging populations. But on the other hand, future benefits are reduced and individuals are exposed to additional risks. These divergent movements create a growing sense of retirement insecurity.

There are, however, options to improve retirement security for an aging population. Demographic trends often do not pose a binding constraint since most countries have sufficient room to increase employment

relative to the working age population, thus growing the tax base for their pension systems. Moreover, reversing the trend towards rising income inequality in many countries could also help to improve the finances of public pensions since they redistribute income from high life time earners to low lifetime earners. Retirement income security can also be improved, in addition to securing public pensions, by designing private pensions in such a way that risks to the individual are minimized.

Retirement Savings

Although the design of retirement systems differs from country to country, there are some similarities. Many countries rely on the government to provide a basic benefit, and individuals are expected to save additional funds for retirement through private pensions or other private savings to ensure a decent standard of living in retirement.

The financing and the benefit distributions of public pensions can schematically be described by equations (1) and (2):

$$I_T = t_T * N_T * y_T \quad (1)$$

where I is total public pension income, t is the combined tax rate in period T , N is the total number of covered employees in period T , and y are the average earnings subject to public pension taxes. Importantly, the total funds available to finance public retirement benefits grow if the number of workers rises and if average earnings increase.

Benefit payments for public retirement benefits can be described by:

$$B_T = \sum b_{iT} \quad (2)$$

$$b_{iT} = \frac{\sum w_{BT}}{TIME} * \rho * \delta * I_{BT}. \quad (2')$$

Total benefits in period T are the sum of individual benefits b_{iT} . Individual benefits are determined by average earnings over individual earnings histories, $(\sum w_{BT, TIME})/TIME$, where $TIME$ is equal to maximum pensionable years. The average

wage is then multiplied by a replacement rate, ρ , by a redistributive factor, δ , and by a benefits indexation factor, I_{BT} .

Public pension benefits increase if the number of beneficiaries rises, if average earnings over the benefit calculation period grow, and if benefits are indexed to rising prices or wages. Hence, benefits can rise because of improving economics and because of rising generosity in the benefits formula, such as a change in indexation.

In many countries, expenditures for public pensions were high compared to other social expenditures, and they have been rising in many places for some time. In 1995, expenditures amounted to more than 10% of GDP in France, Germany and Italy, and to a little over 5% in Japan and the U.S. (OECD, 2000). Moreover, these expenditures have been increasing as a share of GDP for a number of years in many OECD countries. From 1980 to 1995, Germany and the U.S. had the smallest increases with +0.3%, while they were largest in Japan (+2.2%), France (+2.6%), and Italy (+3.6%), and more moderate in Sweden (+1.3%) and the UK (+1.4%) (Weller 2001a).

Tax rates have often increased, too, but not necessarily in tandem with expenditures. Tax rates increased in Germany and the U.S. since 1961; French tax rates grew fastest in the 1970s; Italian tax rates increased most rapidly in the 1990s; Swedish tax rates rose sharply in the 1960s; Japanese tax rates dropped in the 1980s before increasing in the 1990s; and UK tax rates increased in the 1970s, before declining in the 1980s and rising in the 1990s.

The large public expenditures on public retirement benefits are reflected in their relative importance as source of income for retirees. They provided a low 58% of retirees' household income in Italy and a high 83% in West Germany in 1989. More than two thirds of retirees' household income came from

pensions (81%) in France in 1989, in the UK (68%) in 1991, and in the U.S. (67%) in 1991 (Hauser 1998).

In addition to publicly pensions, many countries rely on private pensions for retirement income. Typically, the term private pension refers to an employment based retirement savings vehicle. Other terms include retirement savings plans or superannuations, reflecting the specific design of the pension plan. Savings in a pension constitute part of an employee's compensation. Despite a wide variety of pension design, there are essentially two broad categories: defined contribution and defined benefit plans.

Under a defined benefit plan, the employee is guaranteed a benefit upon retirement, usually based on years of service, age and final earnings. The benefit formula is often designed such that employees accrue most of their benefits during their last years of service. For example, a formula that relates the retirement benefit to the average earnings of the last few years implies that the worker gains most of his or her benefits during those years, when his or her earnings are close to their peak.

Such a pension plan design reflects the desire of employers to use pensions as retention tools. By promising disproportionate benefit accrual in the later years of employment, employees have an incentive to remain with an employer for an extended period of time.

Similarly, employers attempt to raise employee loyalty by requiring that employees have to work for the employer for a minimum period before they become eligible for the pension benefit, or vest in the pension.

Workers in many countries are becoming increasingly less likely to have a defined benefit pension, though. Instead, they have a defined contribution plan. For example, the share of U.S. households with a defined

contribution plan rose from 7.9% in 1983 to 47.8% in 1998, whereas the share of households with a defined benefit plan fell from 67.8% to 45.9% during the same period (Wolff 2002).

A number of countries have made efforts to establish the legal framework to promote more defined contribution plans. France attempted to introduce a new legal framework for private pensions in 1997, albeit unsuccessfully (IBIS 1998). Italy began introducing reform legislation to promote private pensions, especially defined contribution plans, as early as 1993 and continued with further reform efforts in 1995 and 2001 (IBIS 1997,2000a). And Japan completed legislation in 2000 that would facilitate the introduction of defined contribution plans (IBIS 2000b). Other countries, such as Germany or Korea, have either already introduced or are contemplating the introduction of regulations for defined contribution plans.

In a defined contribution plan, the employee bears the risks, not the employer, i.e. there are no guaranteed benefits. Employees, employers, or both contribute a share of the employee's income to an account. The investment choices, vendors, and fees are often controlled by the employer or regulated by the government. Employees bear the risks of their investment's performances, although the employer can make the investment choices. Employee and employer contributions can be pre-income or post-income tax. In countries with progressive income taxation, pre-income tax contributions are a subsidy that increases relative to income with income. Also, while defined benefit plans often pay monthly benefits as long as the retiree is alive, defined contribution plans offer the option to convert savings into monthly benefits, so-called annuitization, or to receive a lump sum distribution or both.

Employees face a number of risks under a defined contribution plan that are borne by the employer under a defined benefit plan. For one, defined contribution plans often do not offer the same insurance protection that public pensions, and many defined benefit plans, offer, since they may not include disability or survivorship benefits. Additional risks include the idiosyncratic risk of making unwise investment choices, and the risk that financial markets generate low rates of return for long periods of time, and thus prevent the employee from saving enough for retirement. Further, because defined contribution plans more often than not offer no option to convert accumulated savings into a lifetime stream of income payments, there is the risk that a retiree will outlive her or his savings. Because of the insurance character of public pensions and defined benefit pensions, retirees do not face the same longevity risk as under a defined contribution plan. Moreover, recent examples in the UK and the U.S. have shown that many new retirement savings vehicles are vulnerable to fraud and deception.

There are also so-called hybrid plans, which combine aspects of defined benefit and defined contribution plans. For example, an employee can receive a minimum guarantee on either the rate of return or on part or all of their assets, plus an additional benefit if his or her investments performed above a certain threshold. For instance, many teachers and college professors in the U.S. participate in TIAA-CREF, which offers both a guaranteed benefit and a defined contribution plan to participating employees. Also, Australian pension plans, so-called superannuation funds, pay benefits as a mixture of guaranteed defined benefits and a variable productivity benefit.

Funding Systems

In recent years, many industrialized countries have emphasized private pensions over public

pensions, occasionally even replacing part of their public pensions with defined contribution plans. For instance, the German government eased the way for defined contribution plans with the Third Financial Markets Promotion Law of 1998 (OECD 2000b) and new pension reform legislation introduced in 2001 (IBIS 2001). The latest changes in German pension law were combined with reductions in scheduled benefits of the public pension system (Weller, 2001a). Further, Sweden changed its public retirement benefit from a defined benefit system to a partial defined contribution system in 1998 (Ministry of Health and Social Affairs 1998). Also, since the 1980s, workers in the UK had to rely increasingly on defined contribution plans for their pension in addition to a shrinking basic public pension (OFT 1997).

The shift from public to private pensions also entails a change in the funding structure of retirement benefits. An important issue that is often raised with respect to public pensions' funding is whether there are unfunded liabilities or whether its liabilities are partially or fully prefunded. The assumption is that prefunded liabilities are less of a burden on future generations than unfunded liabilities.

Many public pensions are financed on a pay-as-you-go basis, whereby current income pays for current benefits i.e., their future liabilities are considered unfunded. Some public pension systems, such as the U.S.' Social Security, hold trust funds that can pay benefits for a number of years, but that still fall short of paying for all future promised benefits. These systems are referred to as partially prefunded. Two-tiered public pension systems – one basic retirement benefit plus an earnings related benefit – such as the Swedish or the UK system have a pre-funding component, as the earnings related public retirement system is pre-funded, but

the basic one is not (Weller 2001a). Replacing part of public pensions with private pensions also means more prefunding as private pensions, especially defined contribution plans, tend to be prefunded.

The distinction between unfunded and prefunded pension liabilities is arbitrary when it comes to public pensions. In either case, future liabilities are covered by a claim on shares of future national income. In the unfunded case, the claim is often solely on future labor income, whereas in a prefunded system the claims are largely on capital income. Since labor and capital income shares have to remain stable in the long-run, there is no advantage in having claims on future shares of one form of income over another.

Although the distinction between unfunded and prefunded public pension liabilities is arbitrary since in both cases beneficiaries have claims on future streams of income, changing from one system to the other has real economic implications. For instance, changing from a largely unfunded pay-as-you-go system to a prefunded individual account system, as was done in Chile in the early 1980s, requires substantial transition costs. In essence, current taxpayers continue to pay for promised benefits accrued under the old system for a period of about 30-40 years. Simultaneously, current taxpayers begin to build up assets in a prefunded system. Thus, at least one generation of taxpayers will pay twice: once for the old system and once for the new one.

Future Challenge for Public Pensions

Changes to public pensions are often justified by an anticipated demographic crisis, i.e. too many beneficiaries for the expected number of taxpayers. The share of over 65-year-olds is expected to almost double in Italy and Japan between 2000 and 2050 (table 2). It is also forecast to grow substantially in other

countries: by 70% in France, by 76% in Germany, by 57% in Sweden, by 71% in the UK, and by 61% in the U.S.

Table 2. 65 Year Olds as Share of Total Population, 1990 TO 2050

	France	Germany	Italy	Japan	Sweden	UK	US
1990	14.02	-	-	11.96	17.79	-	12.50
2000	16.00	16.25	18.09	17.01	17.29	15.67	12.64
2010	16.79	19.70	20.55	21.76	19.18	16.69	13.23
2020	20.61	21.41	23.55	26.83	22.69	19.59	16.52
2030	23.98	25.75	28.15	28.31	25.08	23.50	20.02
2040	26.45	28.43	34.24	31.85	27.11	26.35	20.44
2050	27.25	28.55	36.10	33.86	27.23	26.83	20.30

Notes: All figures are in percent. Source is the U.S. Bureau of the Census, International Data Base.

While expenditures and taxes for public pensions did not rise universally in the past, future increases are predicted almost everywhere. Turner et al. (1998) estimated that public debt relative to GDP would increase to almost 100% in Japan and the EU, and to close to 70% in the U.S. by 2050. Also, Sinn (1999) projected that German tax rates would rise from 20% to over 30% by 2030, and Prognos AG (1998) predicted that German tax rates would rise to 24% by 2040. Further, the trustees of the U.S.' Social Security (SSA, 2002) estimated that expenditures for public pensions as a share of taxable earnings would rise from 11% in 2002 to 23% in 2050. Similarly, the OECD (2000) forecast that public pension expenditures in Italy would rise from about 14.5% of GDP in 2000 to 15.8% in 2032 before falling again to 14.5% in 2050.

However, many forecasts either held economic factors, which may have beneficial effects, constant or ignored them in discussing policy options. For instance, Turner et al. (1998) estimated that living standards in the EU would be 43 percentage

points greater in 2050 with higher employment. Sinn (1999) and Prognos AG (1998) did not consider changes in labor force participation rates or productivity and wage growth in their forecasts (Sinn and Thum 1999). Lastly, the trustees of the U.S. social

security (SSA 2002) assumed in their baseline scenario that low productivity growth will remain unchanged. In comparison, Weller (2004) showed, based on simulations for France, Germany, Italy, Japan, Sweden,

the UK, and the U.S. that stronger employment growth could be useful everywhere, and that faster productivity growth could be useful in every country, except in Germany and Japan due to their indexation of benefits to wages.

Macro Economic Trends and Public Pension Funding

Although most of the discussion about public pensions centers on the demographic gap – more older people, fewer young people – the real constraints to pensions arise from economic trends as equations (1) and (2) show. In particular, employment can grow faster than the working age population, if the employment to population ratio is relatively low. Moreover, productivity gains, if they translate into wage increases, can allow fewer workers to pay for more beneficiaries (Weller 2001a). During the 1980s and 1990s, though, economic trends helped to worsen the funding outlook for public pensions as employment and wage growth slowed.

Faster employment and productivity growth, as well as declining earnings inequality could help to improve public pension finances (Weller 2003). Lower

employment growth means fewer taxpayers and slower wage growth means slower growth of the taxable payroll (equation (1)). Alternatively, since benefits are linked to earnings histories, higher wages in the past will translate into higher benefits in the future (equation (2)). But during the transition period from a low wage to a high wage regime public pension funding should improve. If wage growth falls below productivity growth, income is redistributed towards capital, thereby reducing contributions below their potential, and vice versa. Also, where a cap on income exists above which contributions are not collected, less earnings inequality means that fewer total earnings fall beyond the cap leaving a larger tax base as a result (SSA 2002).

Often employment growth slowed in the OECD since the 1970s. Japan, Sweden and the U.S. had average employment to working age population ratios above 70% in the 1990s. But lower employment to population rates persisted in the UK with less than 70%, in Germany with 65%, in France with less than 60%, and in Italy with 52% in the 1980s and 1990s. Further, in some cases of high employment levels, they continued to grow (U.S. and Japan), or stayed high (Sweden). However, in France, Germany, Italy, and the UK employment relative to the working age population declined steadily since the 1960s.

Lower employment partially followed early retirement options. From 1973 to 1982—when the normal retirement age was lowered to 60—a general system was in place in France intended to provide 60-70% of income to workers who had lost their job at age 60 or above (Blanchet & Pelé 1997; Holcblat et al. 1999). In Germany, about 2.3 million out of 5.9 million retirees received early retirement pensions, e.g. women were entitled to full benefits at age 60 instead of 65 as it was for men. And long-term unemployed workers also have the option to retire early

(BfA, 2000). In Italy, workers could retire early with full benefits after 35 years of service (Brugiavini, 1997).

Moreover, wage growth slowed as productivity growth slowed and wage growth fell below productivity growth in the 1980s and 1990s (Weller 2003). Slower wage growth and a shift from labor income to capital income reduced the tax base.

Also, the distribution of income within labor became more unequal as earnings inequality grew in many countries in the 1980s and 1990s. Gottschalk and Schmeeding (1997) found that inequality grew fastest in the UK and in the US, and the least in the Nordic countries. Other studies supported some of these findings. Several studies indicated rising earnings inequality in France between 1976 and 1987 (Katz et al 1995). Japan's earnings inequality grew between 1974 and 1990 (Katz et al 1995). Sweden's earnings inequality rose during the 1980s, with stronger growth of inequality in the second half of the 1980s than before (Edin and Holmlund, 1995). Moreover, in the UK wage inequality rose at double digit rates between 1979 and to 1990 (Freeman & Katz 1995; Katz et al 1995). Earnings inequality grew sharply in the U.S. between 1979 and 1990 (Freeman & Katz 1995). Inequality appeared to remain stable in Germany in the 1980s (Abraham & Houseman 1995; Katz et al 1995). For Italy, Freeman and Katz (1995) reported signs of expanding wage differentials by occupation and education in the late 1980's, while others found that inequality remained the same in the 1980's (Abraham & Houseman 1995; Katz et al 1995).

Worsening public pension finances could also result from rising benefits. These can be caused by greater longevity, benefit improvements, faster indexation, higher wages during the preceding decades and greater inequality. As people live longer, they

receive more benefits. Also, more people may become eligible, e.g. through early retirement benefits. Due to benefit indexation, faster price or wage growth leads to higher benefits. And public pensions tend to redistribute income towards lower lifetime earners. If their number rises due to more inequality, benefits grow disproportionately faster than taxes. However, during the 1980s and even more so during the 1990s, the trend was towards reducing public pension benefits e.g., by increasing the normal retirement age or by changing the benefit indexation factor, than to increase benefits (Weller 2001a).

The figures highlight the importance of economic factors for public pension finances. In France, Germany, and Italy, where tax rates rose in the 1980s and 1990s, employment relative to the working age population was lowest and declining. Also, Germany and Italy had the lowest wage growth rates during that period, and France had wage growth of less than one percent annually (Weller 2001a). Similarly, in Japan, where tax rates rose in the 1990s much faster than they had fallen in the 1980s, wage growth was only half a percent per year during the 1980s and 1990s, and earnings inequality seems to have risen. Further, the UK saw rising tax rates in the 1990s in the face of falling employment and rapidly rising earnings inequality. In comparison, Sweden had stable employment and strong wage growth, and falling or flat tax rates. The fact that U.S. tax rates did not rise despite low wage growth appear to be a result of strong employment growth (Weller 2001a).

Policy Responses to Future Challenges

As the populations in many countries are expected to grow, the policy challenge is to provide adequate retirement income for all elderly. This requires that countries maintain, if not improve existing public pensions, and

that policies are in place to build secure private retirement savings.

Most of the discussions surrounding public pensions have either taken demographic forecasts as given or ignored possibly beneficial effects from stronger employment and wage growth or lower inequality. Subsequently, policymakers have often focused on prefunding their pensions through the promotion of private pensions, sometimes even replacing parts or all of existing public pension systems with private pensions. At the same time, many countries have begun to reduce the generosity of their public pensions e.g., by raising the retirement age (Weller 2001a).

Fewer public pension benefits and the replacement of public pensions with defined contribution plans are inadequate responses to the challenges for an aging workforce. For one, economic research has shown that households will not save enough to compensate for the loss of benefits when the generosity of public pensions is reduced (Weller 2001b).

This is especially true for one particular benefit cut, a higher retirement age. Although longevity has improved, the health of older workers has not necessarily seen similar improvements. Hence, a substantial share of older workers may face the hard choice of working longer in poor health or of retiring without adequate benefits when the retirement age is increased (Weller 2001b).

Further, the replacement of part or all of public pensions with private pensions, especially defined contribution plans, means an increase in the risks for future retirees. Implicit government guarantees that underlie public pensions are replaced with riskier private market promises of uncertain wage growth and rates of return.

The challenge for policy makers in most industrialized countries and many industrializing economies is to provide

adequate retirement savings for an aging population. This requires first and foremost to secure the commitment to providing a public pension as basic retirement income support. Since public pensions are typically insufficient to provide adequate retirement income for everybody, public policy should create an environment, in which individuals can accumulate sufficient retirement savings. This also implies that retirement savings are secure, so that they are available when an individual chooses to retire.

To bolster public pensions, policymakers should focus on raising the employment share of the working age population, and on equalizing the income distribution between capital and labor and within labor. Improving the employment outlook in many OECD countries will require a combination of macro and micro policies. In particular, many OECD countries have found themselves in a deflationary macro economic environment after the boom years of the 1990s. The macro economic environment is characterized by a debt overhang in the U.S., tight monetary policies in the EU, and deflation in Japan. A pro-growth policy response may include, among other things, a sound economic stimulus in the U.S., a loosening of monetary policy in the EU, and possibly inflationary policies in Japan. In addition, policy makers may want to consider micro policies to increase employment, wherever possible. Such micro policies may include training programs for the long-term unemployed, employment transition programs for young adults, or employment support programs, especially for parents.

In addition, policy makers may want to consider public policies that will help to equalize the distribution of income between labor and capital and within labor. Such policies include strong worker rights that will promote unionization and collective bargaining as a regulative on capital's power.

Additionally, public policies can strengthen labor's bargaining position through minimum wage laws, and strong social safety nets. Further, a number of policies appear to help reduce earnings inequalities. Particularly, equal access to quality primary and secondary education seem to be a consistent policy.

In addition to pursuing policies that could help to strengthen public pensions, policy makers should also consider policies that will promote sound private pensions. These include mandatory universal coverage i.e., all employers have to contribute a minimum share of payroll for their employees to a private pension account. This appears particularly pertinent, since an increasing share of the workforce in many industrialized countries have a more loose employment relationship than in the past. However, so-called contingent workers are often less likely to have similar benefits as traditional full time workers. Mandating pension coverage for all employees may thus help to reduce inequities in private pension systems. Systems of voluntary private pension coverage appear to create large inequities, while leaving a large share of workers uncovered (Wolff, 2002). In addition, a number of policy options exist to secure pension savings. Among these policies are mandatory diversification of assets, public insurance of assets, immediate vesting, full portability of pension savings between jobs, financial industry regulation to avoid conflicts of interest, low cost investment options through the government and so on.

Only a combination of strong public pensions and secure private pensions will ensure that industrialized and industrializing countries are adequately prepared to provide retirement income to an aging population.

Selected References

Abraham, K. and S. Houseman. (1995)
"Earnings Inequality in Germany", in R.
Freeman and L. Katz (Editors),

- Differences and Changes in Wage Structures*. Chicago IL: University of Chicago Press.
- Blanchet, D. and L. Pelé. (1997) *Social Security and Retirement in France*. NBER Working Paper No. 6214, Cambridge MA: National Bureau of Economic Research.
- Boeri, T.; A. Börsch-Supan and G. Tabellini. (2000) "Would You Like to Shrink the Welfare State? A Survey of European Citizens", *Economic Policy*, 32, 9-50.
- Brugiavini, A. (1997) *Social Security and Retirement in Italy*. NBER Working Paper No. 6155. Cambridge MA: National Bureau of Economic Research.
- Bundesversicherungsanstalt fuer Angestellte. (BfA) (2000) *Statistik*. www.bfa-berlin.de
- Commission to Strengthen Social Security. (CSSS) (2001) *Strengthening Social Security and Creating Personal Wealth for All Americans*. Washington, D.C.: CSSS, www.csss.gov
- Edin, P. and B. Holmlund. (1995) "The Swedish Wage Structure: The Rise and Fall of Solidarity Wage Policy?", in R. Freeman and L. Katz (Editors), *Differences and Changes in Wage Structures*. Chicago IL: University of Chicago Press.
- Freeman, R. and L. Katz. (1995) (Editors) *Differences and Changes in Wage Structures*. Chicago IL: University of Chicago Press.
- Gottschalk, P. and T.M. Smeeding. (1997) "Cross-national Comparisons of Earnings and Income Inequality", *Journal of Economic Literature*, 35, 2, 633-687.
- Hauser, R. (1998) *Adequacy and Poverty Among the Retired*. OECD Aging Working Paper AWP 3.2. Paris, France: OECD.
- Holcblat, N.; P. Marioni and B. Roguet. (1999) *Politiques D'Emplois Depuis 1973, Données Sociales*. Paris, France: INSEE.
- IBIS News Database. (2001) *The Government's Pension Reform Legislation was Passed by the Lower House of Parliament, Briefing—Pension Reform Passed by Bundestag. Germany*. 02/01, Chicago: Charles D. Spencer and Associates.
- IBIS News Database. (2000a) *The Tax Deductible Limits on Employee and Employer Pension Plan Contributions Have Been Increased, Briefing—Pension Funds—Contribution Limits. Italy*. 06/00. Chicago: Charles D. Spencer and Associates.
- IBIS News Database. (2000b) *Details about the Proposed Legislation to Permit Defined Contribution Plans in Japan are now Becoming Available, Briefing—Occupational Pensions—Defined Contribution Plans. Japan*. 02/00, Chicago: Charles D. Spencer and Associates.
- IBIS News Database. (1998) *The "Loi Thomas" Will Be Replaced, Briefing—Pension Funds—Loi Thomas to be Replaced. France*. 11/98. Chicago: Charles D. Spencer and Associates.
- IBIS News Database. (1997) *Occupational Pension Law, Briefing—Pension Law Takes Effect—Summary of Provisions. Italy*. 9/97. Chicago: Charles D. Spencer and Associates.
- Office of Fair Trading. (1997) *Report of the Director General's Inquiry into Pensions*. London UK: Office of Fair Trading.
- Organization for Economic Cooperation and Development. (OECD) (2000) *OECD Economic Surveys: Italy*. Paris, France: OECD.
- Organization for Economic Cooperation and Development. (OECD) (2000b) *Germany: Questionnaire 2000*. Original Report, January 27. www.oecd.org/subject/ageing
- Organization for Economic Cooperation and Development. (OECD) (2000c) *Social*

- Expenditure Database (SOCX)*. Paris, France: OECD.
- Prognos A.G. (1998) *Auswirkungen veränderter ökonomischer und rechtlicher Rahmenbedingungen auf die gesetzliche Rentenversicherung in Deutschland*. DRV-Schriften, Band 9, Frankfurt am Main.
- Sinn, H. (1999) *The Crisis in Germany's Pension Insurance System and How It Can Be Solved*. NBER Working Paper No. 7304. Cambridge MA: National Bureau of Economic Research.
- Turner, D.; C. Giorno; A. De Serres; A. Vourch and P. Richardson. (1998) *The Macroeconomic Implications of Ageing in a Global Context*. OECD Ageing Working Paper AWP 1.2, Paris, France: OECD.
- U.S. Bureau of the Census. (2002) *International Data Base*. Washington, D.C.: Census.
- U.S. Social Security Administration. (SSA) (Various Years) *Social Security Programs Throughout the World*. Washington DC: SSA.
- U.S. Social Security Administration. (SSA) (2002) *The 2002 Annual Report of the Board of Trustees of the Federal Old-Age and Survivors Insurance and Disability Insurance Trust Funds*. Washington, D.C: SSA.
- Weller, C. (2004) "The Future of Public Pensions in the OECD", *Cambridge Journal of Economics*, 28, 4, 489-504.
- Weller, C. (2001a) "Programs without Alternatives: Public Pensions in the OECD", in D. Jacobs and S. Friedmann (Editors), *The Future of the Safety Net: Social Insurance and Employee Benefits in the 21st Century*. IRRRA Research. Ithaca NY: Cornell University Press.
- Weller, C. (2001b) *Can Workers Afford a Higher Retirement Age?* EPI Technical Paper No. 255. Washington, D.C.: Economic Policy Institute.
- Wolff, E. (2002) *Retirement Insecurity: The Income Shortfalls Awaiting the Soon-to-Retire*. Washington, D.C.: Economic Policy Institute.

Christian E. Weller
 Center for American Progress
 Washington D.C., USA.
 cweller@americanprogress.org

Pharmacoeconomics

Edward J. O'Boyle

Introduction

Modern pharmaceuticals—prescription and over-the-counter (OTC) medicines—begin with the introduction of aspirin in 1899 by Bayer & Company. The very same Bayer chemist who synthesized aspirin synthesized the morphine derivative which was named heroin because it made the workers on whom it was tested feel heroic. Mistakenly, heroin was thought to be non-addictive and was used to treat coughs, pain of childbirth, war injuries, and mental disorders before it was banned in most countries in the 1930s (Chemical Heritage 2001).

One hundred years after it was introduced annual aspirin use worldwide exceeded 100 billion tablets for treatment of a variety of ailments including fever, migraine, rheumatoid arthritis, and acute tonsillitis (Aspirin Foundation (a)(b)). Penicillin, the first antibiotic to be isolated, was used clinically for the first time in 1941. Today the World Health Organization identifies a total of 323 drugs as essential and uses that list to define and measure the access which a country's population has to essential medicines (WHO 2005:112-119).

According to WHO smallpox as a human disease was eradicated on a global basis in 1979, though it could be re-introduced by bio-terrorists. In 1988 WHO resolved to eradicate polio from the world. That objective has not yet been met. WHO is wary that once the virus is eradicated it might re-emerge due to an accidental or deliberate release (WHO 2007:xviii, xxi).

Sales, Profit and Public Benefit

Worldwide pharmaceutical sales amounted to an estimated \$690 billion in 2007 (IMS 2007:1). At the same time, prescription sales in the United States amounted to \$286 billion. The three leading drug classes by sales were lipid regulators (\$18.4 billion), proton pump inhibitors (\$14.1 billion), and antipsychotics (\$13.1 billion) (IMS 2008:1).

The three leading drugs sold worldwide in 2004 were Pfizer's Lipitor with sales of \$10.9 billion, Merck's Zocor with sales of \$5.2 billion, and GlaxoSmithKline's Seretide/Avadair with \$4.5 billion in sales. Each one of 50 different drugs in 2004 had sales worldwide in excess of \$1 billion (Davidson:16, 26).

The cost and price of drugs are issues because together they determine the amount of the manufacturer's gain which is necessary (lower limit) and the amount justified (upper limit). A recent report on the financial performance of 15 major pharmaceutical companies in the world indicated that profit margins in 2004 ranged from a low of -24.0 percent at Sanofi-Aventis and -11.4 percent at Schering-Plough to a high of 26.3 percent for Takeda and 25.3 percent for Merck. Single-year performance measures, however, can be misleading. Sanofi-Aventis in 2003 reported a 25.8 percent profit margin and Schering-Plough in 2002 reported a -19.4 percent profit margin (Davidson:17, 27).

The market by itself cannot resolve the issue of drugs which are priced high enough to retrieve the cost of research and development but are unaffordable for those persons who need those drugs. The problem becomes more ethically complicated when the drugs in question are life-saving. Differential pricing to take account of need and affordability introduces the problems of arbitrage: the reselling of drugs earmarked for poor countries, where they have been priced

to make them more affordable, in rich countries where the higher prices for the same drugs provide opportunities for profiteering. To sustain such price differentials it is necessary for rich countries to forego importing the lower-price drugs from poor countries or for payers and drug companies to adopt such practices as negotiated contracts with confidential rebates (Danzon and Towse 2003:201).

What some may call price differentiation, others may regard as price discrimination. In the United States, for example, it has been asserted that prescription drugs are subject to price discrimination in which senior citizens are charged more than other customers such as HMOs, the federal government, and large corporations.

A report prepared in 2000 by the Joint Economic Committee of the U.S. Congress argued that if just 10 percent of the net economic gains from increased life expectancy can be attributed to publicly funded research, the payoff from the taxpayers' investment in NIH sponsored research is 15:1 (Joint Economic Committee 2000:ii).

Access, Quality, and Compliance

More than four billion persons have regular access to WHO's essential drugs including 12 antiretroviral medicines for the prevention and treatment of HIV/AIDS. 2.1 billion, however, do not have access to essential medicines at affordable prices and assured quality. The critical factors in this access gap are irrational use of medicines, unfair financing, unreliable delivery systems, and unaffordable prices (WHO 2003).

Fewer than one-third of developing countries have fully functioning drug regulatory agencies. Ten to 20 percent of sampled drugs fail quality control tests in many developing countries. Poor manufacturing practices often result in toxic,

sometimes lethal, products. On a global basis, about one-half of all patients take their medicines correctly. About 75 percent of antibiotics are prescribed inappropriately (WHO 2003).

Regulatory agencies are just becoming aware of the problem of counterfeit drugs and the risk they pose to public health. The first study to compile information on the extent of this problem was published in 2003. Addressing this problem will require more regulatory oversight in which counterfeit goods are seen as a disease mechanism (see Forzley 2003:ii,30).

Research and Development

An estimated \$32 billion was spent in 2002 on R&D by pharmaceutical companies worldwide. Of that, \$26.4 million was spent in the United States. For PhRMA (Pharmaceutical Research and Manufacturers of America) members, R&D expenditures represent 18.2 percent of domestic sales. In nominal dollars, spending today on research and development worldwide is more than 25 times greater than the \$1.3 billion spent in 1977 (PhRMA 2003a:10).

The U.S. National Institutes of Health in fiscal year 2001 allocated approximately \$16.2 billion to support basic (extramural) research at universities, medical centers, hospitals, and research institutions, and about \$2.0 billion for the NIH's own (intramural) research labs. This funding is intended to support the NIH's mission to "sponsor and conduct medical research and research training that expands fundamental knowledge about the nature and behavior of living systems; improves and develops new strategies for the diagnosis, treatment, and prevention of disease; reduces the burdens of disease and disability; and assures a continuing cadre of outstanding scientists for future advances" (NIH 2001:3).

In fiscal year 2004 NIH has been budgeted \$22.2 billion for extramural research and \$2.7 billion for intramural research. Support for research needed in the war on terrorism was first priority for program increases (U.S. Dept of Health and Human Services 2004:31).

In April 2007, PhRMA reported that biopharmaceutical researchers were testing 219 medicines to meet the needs of children including children with cancer and genetic, neurologic, and respiratory disorders (PhRMA 2007a:1). Later in 2007, PhRMA reported that there were more than 700 medicines in development for diseases that affect only or disproportionately women including women suffering from cancer, arthritis and musculoskeletal disorders, and autoimmune disorders. Research and development expenditures across the pharmaceutical industry in 2006 totalled \$55.2 billion (PhRMA 2007b:1-2)

Just one in every 100 research projects results in a new drug, the development process lasts on average 14 years, and it takes six to seven years to recover the research and development costs once the drug is available for sale to the public (PhRMA no date: video) Tufts Center for the Study of Drug Development estimated that in 2006 the fully capitalized cost for developing a new prescription drug averaged \$1.2 billion (Tufts 2006:1).

Does the cost of research and development justify the price of new prescription drugs? PhRMA says that it does, and denies that government funds play a major role in research and development. In a 2001 Report to Congress, NIH specifically refuted the misconception that the government pays for most of the research on top-selling prescription drugs (PhRMA 2003a:11).

Families USA argued that drug prices can be reduced without suppressing research and development because: (1) the financial future of any pharmaceutical firm depends on its

development of new, successful products and therefore it cannot cut back on research and development; and (2) retail prices reflect large outlays on advertising, marketing, and administration which can and should be cut back (Families USA 2001:10). Another critic claimed that the income tax credits awarded to U.S. drug companies are used to increase dividends and not to reduce prices (New York 1998:1).

A World Bank discussion paper states that the low return on research and development expenditures is a barrier to the development of medicines needed to fight diseases in developing countries. Between 1975 and 1997 only 13 new products were approved specifically for tropical diseases (Govindaraj 2003:13). Local production is a solution only if the cost of building local production capacity is not excessive, product quality is assured, and pricing is competitive with products from existing foreign generic manufacturers (Kaplan & Liang 2005).

NIH calls attention to the special problems in sorting through the linkages between government expenditures on basic research and new drug development:

“Analysis of the 47 therapeutic drugs that have reached annual sales in the U.S. of \$500 million, and determination of which of these had intellectual property that ties back to federal funding, was particularly difficult ... due to the fact that (federal) regulations do not require that investigators provide such information to the funding agency, and ... tracking down the “pedigree” of these drugs has to be done manually on a case-by-case basis. ... (It) is not possible to cross-reference NIH grants and contracts that funded inventions with any patents or licenses embodied in the final product. Nor is it possible to identify other federal and/or non-federal sources of funds that contributed to an inventive technology” (NIH 2001:9).

Outcomes Research

Frank Lichtenberg is one of the most widely cited researchers on the benefits and costs of pharmaceuticals, especially new prescription medicines. He is the holder of an endowed chair in economics at Columbia University and a research associate at the National Bureau of Economic Research. His work has been cited approvingly by the U.S. Food and Drug Administration (2003), Heritage Foundation (2002), PhRMA (2003b,c), Irish Pharmaceutical Healthcare Association (2002), National Pharmaceutical Council (2002), Cato Institute (2003), Institute of Medicine (2002) and by others including numerous professional colleagues. His standing and reputation have been enhanced by the 1998 Schumpeter Prize and the 2003 Milken Institute Award for Distinguished Economic Research. We refer below to five of his more influential studies.

In a 1998 study Lichtenberg claimed that “there was a highly significant positive relationship across diseases between the increase in mean age at death and the share of new drugs (that represent an advance over available therapy) in total drugs prescribed by doctors” (Lichtenberg 1998:12). Lichtenberg admits that the data referred only to prescriptions written during office visits, that no account is taken of mistakes made in prescribing drugs or of noncompliance by the persons taking the drugs (Lichtenberg 1998:8). However, nowhere in his econometric modeling does he account for changes in human capital among the health-care providers treating the same patients along dimensions other than drug therapy during the two time periods (1970-80 and 1980-91) covered in his study.

In an earlier study, Lichtenberg reported that “The estimates imply that an increase of 100 prescriptions is associated with 16.3 fewer hospital days. A \$1 dollar increase in

pharmaceutical expenditure is associated with a \$3.65 reduction in hospital care expenditure ... but it may also be associated with a \$1.54 increase in expenditure on ambulatory care” (Lichtenberg 1996:388). In this study Lichtenberg leaves aside a measure of the change in human capital of health-care providers over the 1980-91 period covered.

In a 2002 study, Lichtenberg asserted that given the increase in the age of death between 1979 and 1998, which he attributed to the use of priority drugs the rate of return on pharmaceutical research and development, is 18 percent. Lichtenberg acknowledges that the introduction of new medical devices such as stents and artificial hearts during the study period may bias his results (Lichtenberg 2002a:20,25), but does not include changes in human capital such as occur when younger more recently trained physicians replace older physicians, when major improvements are made in the physician’s knowledge base such as in oncology, emergency medicine, and neonatology, when a physician is board-certified in his/her specialty, and when continuing education requirements to maintain licensure are strengthened.

In another study published in 2002, Lichtenberg reported that based on data for 1996-98 an \$18 increase in new drug expenditures leads to a \$129 decrease in other health-care expenditures, mostly hospital costs (Lichtenberg 2002b:1). Lichtenberg makes no allowance for differences in the hospitals to which the subjects were admitted where costs vary by type (e.g., community, proprietary, university, charity), for differences in the physicians where fees vary by type (e.g., neurosurgeon, obstetrician, urologist), or for differences in payers where early discharge may vary by type (e.g., private insurance, HMO, Medicaid, Medicare, no pay). Miller and Frech are critical of some of Lichtenberg’s work for several reasons

including the omission of relevant variables (Miller & Frech 2002:18).

In none of these articles does Lichtenberg acknowledge that some of the newer drugs were antibiotics which were developed specifically as replacements for earlier antibiotics that through repeated use had become ineffective against drug-resistant viruses, or for differences among patients regarding allergic reactions. Two of these articles are cited by PhRMA, though in both instances inaccurately. With regard to the second article, PhRMA cites the \$3.65 decrease in hospital care expenditures without mentioning the \$1.54 increase in ambulatory care expenditures associated with a \$1 increase in pharmaceutical expenditures (see PhRMA 2001:2; Lichtenberg 1996:388). Regarding the fourth article, PhRMA claims that Lichtenberg compared drugs which were 15 years newer than other drugs when in fact the difference was 9.5 years (see PhRMA 2003a:31; Lichtenberg 2002b:5-6).

In a 2001 article which appeared in *Health Affairs*, Lichtenberg reports that the use of newer drugs reduces mortality, morbidity, and total medical spending (Lichtenberg 2001:250). This article is significant because in it he admits to receiving financial support from the National Pharmaceutical Council which is funded by more than 20 U.S. research-based pharmaceutical companies. His curriculum vitae indicates that he has been supported by 20 private organizations and public agencies including the American Enterprise Institute, Merck, and Pfizer, in addition to the National Pharmaceutical Council and that he is a director of LECG which represents itself as a "leading expert services firm" and includes health care as one of its domains of expertise (Lichtenberg 2003:1,9). Some of his professional colleagues who cited his work approvingly also have been supported by the

pharmaceutical industry (see, for example, Shaw et al 2002:24).

Implications

First, there is a need for better-informed research contributions and suggestions from the economics profession, lest more harm than good is done. Lichtenberg's precise estimates are misleading because they are derived from econometric models which omit critical variables, and suggest a level of confidence in his findings which is misplaced.

Nobel laureate Gary Becker recently scolded the FDA for driving up drug prices by holding unreasonably to a standard of efficacy. Becker recommends using only the safety standard which the FDA applied prior to 1962 (Becker 2002:16). He would have physicians prescribe drugs for which there is no proof that they do any good, only the assurance that they do no harm. Becker's recommendation in effect renders patients into subjects fit for prescription medication by trial-and-error and physicians into targets for malpractice lawsuits. His suggestion could delay the utilization of the most effective therapy and in the extreme could contribute to morbidity and mortality.

Second, Lichtenberg's ties to major pharmaceutical manufacturers which have huge investments in research and development to defend and justify, especially given the prices they charge in order to recover their research and development costs, makes one question the objectivity of his work and whether he insists that it be cited fully and carefully by his backers. The financial stakes are simply too great to rely exclusively or even primarily on PhRMA and PhRMA-supported research for outcomes research. Longitudinal studies which generate micro data and which are carefully specified, as is the case with FDA trials to determine safety and efficacy, are necessary for

authentic pharmacoeconomic research. This is no simple task. WHO identifies ten risk factors contributing to morbidity and mortality: underweight and overweight, unsafe sex and water, high blood pressure and cholesterol, tobacco and alcohol consumption, sanitation and hygiene, iron deficiency, and indoor smoke from solid fuels (WHO 2002a:7).

Third, financial gain is essential if private companies are to engage in pharmaceutical research and development. If reasonable gain leads to drug prices which are unaffordable, those who can afford the drugs must help make them more affordable through price-discrimination schemes or governments must subsidize either the manufacturer or user. If it becomes necessary for governments to assume the research and development role, it would be better to support several research labs rather than one on grounds that the scientific method depends on verifiability from independent sources.

Finally, tuberculosis, malaria, and HIV/AIDS annually claim approximately six million lives worldwide (Global Fund:6). All three are intertwined with poverty. Tuberculosis and malaria long have been treated successfully with drugs, but both present problems of drug resistance. HIV/AIDS, on the other hand, claims more lives than the other two, and is treated today with antiviral therapies which suppress HIV and circumvent the problem of drug resistance but do not cure the disease (CDC 1998:2). Malaria differs from the other two because it is transmitted by the mosquito making prevention difficult. Tuberculosis and HIV/AIDS are transmitted by human contact making prevention much less complicated.

A year's supply of condoms costs \$14 or roughly one-tenth of what it will cost to provide a commonly used triple drug therapy combination to developing countries in Africa and the Caribbean through an October 2003

agreement that cuts the cost of generic antiretroviral medicines in those countries by one-third to one-half. Treating HIV/AIDS properly presents several other problems. First, women who are infected may present with no symptoms of the infection. Second, the tests to confirm the presence of the infection are expensive. Third, the schedule for administering the drugs which are most effective is complicated. Finally, the required medical back-up systems including laboratories to perform the required analysis of the blood samples drawn may not be available in developing countries where the risk of contracting the infection is greatest (Africa-America 2003:1; PhRMA 2001:2; WHO 2002b:28-29). For various reasons, all three killer diseases likely will present major challenges for pharmaceutical research and development and pharmacoeconomics for years to come.

Selected References

- Africa-America Institute. (2003) *Easier Access to AIDS Drugs*. October 23.
- Aspirin Foundation. (a) "The Synthesis of Aspirin". Aspirin Foundation.
- Aspirin Foundation. (b) "100 Years". New York: Aspirin Foundation.
- Becker, Gary. (2002) "Get the FDA Out of the Way, and Drug Prices Will Drop", *Business Week*, September 16.
- Cato Institute. (2003) "Political Assault Against Drugmakers is Misguided, Ignores Benefits", May 8. www.cato.org/new/05-03/05-08-03r.html
- CDC. (1998) "Report of the NIH Panel to Define Principles of Therapy of HIV Infection", April.
- Chemical Heritage Foundation. (2001) "Aspirin Intrigue". www.chemheritage.org/EducationalServices/pharm/asp/asp80.htm.
- Danzon, Patricia and Adrian Towse. (2003)

- “Differential Pricing for Pharmaceuticals: Reconciling Access, R&D and Patents”, *International Journal of Health Care Finance and Economics*, Volume 3, pp. 183-205.
- Davidson, Larry and Gennadiy Greblov. (2005) “The Pharmaceutical Industry in the Global Economy,” summer www.bus.indiana.edu/davidso/lifesciences/lisresearchpapers/pharmaceutical%industry12aug.doc.
- Families USA. (2001) “Off the Charts: Pay, Profits and Spending by Drug Companies”, July. www.familiesusa.org/site/DocServer/drugceos.pdf?docID=767
- Forzley, Michele. (2003) *Counterfeit Goods and the Public’s Health and Safety*. Washington DC: International Intellectual Property Institute, July.
- Global Fund. (2005) “HIV/AIDS, Tuberculosis and Malaria: The Status and Impact of the Three Diseases”, www.theglobalfund.org/en/files/about/replenishment/disease_report_en.pdf.
- Govindarag, Ramesh, Michael Reich, and Jillian Cohen. (2003) *World Bank Pharmaceuticals*. HNP discussion paper, September. *Health Affairs*. www.jnj.com/our_company/healthcare_issues/Health_Affairs_Presents.htm. Johnson and Johnson.
- Heritage Foundation. (2002) *What Seniors Should Know about Government Restrictions on Prescription Drugs*. November 4. www.heritage.org/Research/HealthCare/bg1611.cfm
- IMS. (2008) “IMS Health Reports U.S. Prescription Sales Grew 3.8 Percent in 2007, to \$285.6 Billion”, www.imshealth.com/ims/portal/front/articleC/0,2777,6599_3665_83470499,00.html.
- IMS. (2007) “IMS Predicts 5 to 6 Percent Growth for Global Pharmaceutical Market in 2008, According to Analyst Forecast”, www.imshealth.com/ims/portal/front/articleC/0,2777,6599_3665_82713022,00.html.
- Institute of Medicine. (2002) “Preface”, *Medical Innovation in the Changing Healthcare Marketplace: Conference Summary*.
- Irish Pharmaceutical Healthcare Association. (2002) *Submission to the Commission on Financial Management and Control Systems in the Health Services*. Dublin: July.
- Joint Economic Committee, U.S. Congress. (2000) *The Benefits of Medical Research and the Role of the NIH*, May 2000. jec.senate.gov
- Kaplan, Warren and Richard Liang. (2005) *Local Production of Pharmaceuticals: Industrial policy and Accesses to Medicines*. HNP Discussion Paper. Washington DC: World Bank.
- Lichtenberg, Frank. (1996) “Do (More and Better) Drugs Keep People Out of Hospitals?”, *American Economic Review*, Volume 86, Number 2, pp. 384-88.
- Lichtenberg, Frank. (1998) *Pharmaceutical Innovation as a Process of Creative Destruction*.
- Lichtenberg, Frank. (2001) “Are the Benefits of Newer Drugs Worth Their Cost? Evidence from the 1996 MEPS”, *Health Affairs*, Volume 20, Number 5, pp. 241-51.
- Lichtenberg, Frank. (2002a) *Pharmaceutical Knowledge—Capital Accumulation and Longevity*. Presented at the NBER conference on Measuring Capital in a New Economy, April. www.upenn.edu/ldi/pharmaceutical.pdf
- Lichtenberg, Frank. (2002b) *Benefits and Costs of Newer Drugs: An Update*. NBER Working Paper 8996.

- phrma.org/publications/publications/2002-10-07.584.pdf
- Lichtenberg, Frank. (2003) *Curriculum Vitae*. September. www.gsb.columbia.edu/faculty/flichtenberg/cv.html
- Miller, Richard and H.E. Frech. (2002) *The Productivity of Health Care and Pharmaceuticals: Quality of Life, Cause of Death and the Role of Obesity*. July. repositories.cdlib.org/ucsbecon/dwp/12-02
- National Institutes of Health. (2001) *A Plan to Ensure Taxpayers' Interests are Protected*. July. www.nih.gov/news/070101wyden.htm
- National Pharmaceutical Council. (2002) *Assessing the Impact of Pharmaceutical Innovation*. Washington DC: NPC Health Focus. January.
- New York State Wide Senior Action Council. (1998) *Don't Be Fooled by the Pharmaceutical Industry's Propaganda about Research and Development*. 1998. www.nysenior.org/Issues/Prescriptions/research.html
- PhRMA (2001) *The Value of Medicines*. www.phrma.org/publications/publications/value2001/value2001.pdf
- PhRMA. (2003a) *Pharmaceutical Industry Profile 2003*. Washington D.C.
- PhRMA. (2003b) *Pharmaceuticals are Substituting for More Costly Health Services Like Hospital and Nursing Care*. www.phrma.org/actions/printFriendlyPage.cfm?t=46&r=785
- PhRMA. (2003c) "Rx Spending Growth Slows to Lowest Level in Six Years", *New Medicines, New Hope*, pp. 1-2.
- PhRMA. (2007a) "More Than 200 Medicines in Testing to Meet the Needs of Children", www.phrma.org/news_room/press_releases/more_than_200_medicines_in_testing_to_meet...
- PhRMA. (2007a) "More Than 700 Medicines Now in Development for Major Diseases Affecting Women", www.phrma.org/news_room/press_releases/more_than_700_medicines_now_in_development...
- PhRMA. (no date) "From Molecules to Medicine", video. www.phrma.org/medicines_in_development
- Shaw, James, William Horrace, and Ronald Vogel. (2002) *The Productivity of Pharmaceuticals in Improving Health: An Analysis of the OECD Health Data*. Mimeo. Syracuse University: Economics Department.
- Tufts Center for Study of Drug Development. (2006) *Average Cost to Develop a New Biotechnology Product is \$1.2 Billion*. November. <http://csdd.tufts.edu/NewsEvents/NewsArticle.asp?newsid=69>
- U.S. Department of Health and Human Services. (2004) FY2004 Budget in Brief. hhs.gov/budget/04budget/fy2004bib.pdf
- U.S. Food and Drug Administration. (2003) *Technology and Innovation: Their Effects on Cost Growth of Healthcare*. Statement of FDA Commissioner Mark McClellan before the Joint Economic Committee, U.S. Congress, July 9. www.fda.gov/ola/2003/healthcare0709.html
- World Health Organization. (2002a) *World Health Report 2002*. www.who.int/whr/en
- World Health Organization. (2002b) *Report on Infectious Diseases*. www.who.int/infectious-disease-report
- World Health Organization (2003). *Essential Drugs and Medicines Policy*. www.who.int/medicines/rationale.shtml
- World Health Organization. (2005) *The Selection and Use of Essential Medicines*. www.who.int/medicines/services/expertcommittees/essentialmedicines/TR5933SelectionUseEM.pdf

World Health Organization. (2007) *Annual Report 2007*.

www.who.int/whr/2007/whr07_en.pdf

Edward J. O'Boyle
Mayo Research Institute
West Monroe, Louisiana, USA
edoboyle@earthlink.net

Pollution Rights, Taxes, Permits and Vouchers

Jack Reardon

Introduction

During the twentieth century, markets demonstrated a considerable wealth-producing ability. Thus, it is not surprising that into the twenty-first century, market-based solutions have been advocated for a wide range of social and ecological problems.

Advocates claim that market-based solutions such as taxes, subsidies and pollution permits, are “powerful tools” (Roodman 1997:6) since they utilize the market’s intrinsic advantages and are more flexible and efficient than government-imposed regulations.

Critics, however, maintain that market-based solutions are inappropriate since the market itself is the cause of the problem and only the state can incorporate values of individuals without the ability to pay and of future generations—two constituencies largely ignored by markets.

Nevertheless, efficacious solutions to the global problems of the 21st century, especially global warming, which could potentially “dwarf all other issues” (Flannery 2005:8) will require a judicious mix of market and state policies, along with unprecedented trust, cooperation and communication.

This essay will discuss the merits and shortcomings of market-based environmental solutions and government-imposed regulations. After briefly discussing the relationship between markets, economics and the environment, this essay will contrast the merits and shortcomings of government regulations, taxes and subsidies, and pollution permits, respectively. The final section will offer some concluding observations.

Markets, Economics and Environment

Writing against a backdrop of “empty land, shoals of undisturbed fish, vast forests, and a robust ozone shield”, the 19th founders of neoclassical economics “thought, wrote, and prescribed as if nature did not” (McNeil 2000:335-6). Despite early warning signs of ecological damage, neoclassical economics ignored the social cost of environmental degradation, thus freeing the firm from any responsibility. Pollution was treated as an externality, a “side effect of activities that provide good things” (Krugman 2005:476).

A negative externality occurs when an involuntary cost is imposed on a third party from a voluntary market transaction. A classic example is a firm dumping pollutants into a river, affecting downstream residents. The good is underpriced since the firm does not pay the cost of environmental damage and will result in overproduction.

An externality can be positive, if a third party receives an involuntary benefit from a voluntary transaction. An example is education which benefits the individual and society at large.

Using the word ‘externality’ to describe “pervasive phenomena” (Vatn 2005:261) strips emotional content from the analysis and precludes moral discussion (Cato and Kennett 1999:2-3). Nevertheless neoclassical economics recognized that externalities are market imperfections, compromising the price system’s ability to convey accurate information.

To correct a negative externality, Arthur. Pigou (1920) advocated a tax equal to the inflicted marginal damage. This ‘Pigovian’ tax, as it became known, increases a good’s final price by the marginal cost of polluting activities, sending a message that pollution has a price.

If the tax is set at the ‘correct’ level, “a pareto optimum exploitation of nature, pollution and depletion of natural resources in

accordance with preferences of economic agents is considered possible” (Dietz & Van der Straaten 1992:29).

An alternative to taxation was suggested by Ronald Coase (1960), who conceptualized externalities as ill-defined property rights. By extending property rights to all affected participants, the externality will be corrected.

A property right exists, “between the rights holder and the rights regards under a specific authority like the state granting legitimacy and security” (Vant 2005:254).

Although the concept of property differs across cultures, the formalization of property rights is a western construct (Vant 2005:259). Currently, western society favors private ownership (rather than state, common or open-access) which gives exclusive access to owners while concomitantly excluding non-owners. The state is necessary to define and enforce property rights and to mediate a perpetual “tension or contradiction between access and exclusion” (Singer 1998:245).

According to the Coase Theorem, if the perpetrator of the externality can be identified and property rights are extended to all participants, then the participants can negotiate a proper balance between emission and abatement, regardless of the initial income distribution. The state need only establish and protect property rights and enforce the agreement.

Needless to say, the Coase Theorem “has made a tremendous impact not least on those economists who were against state regulations” (Vant 2005:240). However, the underlying assumptions of perfect competition and zero transaction costs are seldom met, rendering “the theorem of limited relevance to most of the major pollution problems” (Cropper & Oates 1992:680). And the initial income distribution does matter since it determines power relationships, a concept consistently overlooked by mainstream economics.

Command and Control Policies

Environmental degradation and concern for its effects have existed since ancient times (Hughes 1975). Nevertheless, since the Industrial Revolution, government intervention on behalf of the economy was minimal until the middle of the twentieth century.

Instrumental in galvanizing environmental concern was Rachel Carson’s *Silent Spring* (1960) which called attention to the detrimental and cumulative effect of chemicals, many of which were previously assumed benign. In addition, rising incomes in western nations increased demand for a cleaner environment and democracy movements challenged the heretofore carte blanche of business to pollute.

Western nations responded with state-imposed uniform standards. Such standards specified either the technology or an emission rate to combat a specific pollutant for all firms and with a margin of safety sufficiently high so that no adverse health effects would be suffered by any member of the population. They are most effective for either a small number of diffuse polluting sites, or an urgent problem requiring imminent action. They are least appropriate for dealing with a diffuse source of environmental impacts and a large number of smaller companies (ACCA 2003:14).

The Montreal Protocol (1987) which banned the production of ozone depleting substances is considered the most successful example. Given the impossibility of internalizing the damage to humanity of ozone-layer depletion, market measures were eschewed in favor of direct regulations (Anderson & Sarma 2002:351). Nevertheless, the Protocol allowed signatory states the means of implementation, including market measures such as taxes, subsidies and permits.

Another example, although widely criticized by mainstream economists, is the United States Clean Air Act (1969) which mandated technology-based standards requiring uniform emission reductions for plants/facilities within the same industry or age. The initial Act did not require cost-benefit analysis for the individual firm, although economic costs and dislocations to local economies were considered relative to the industry category as a whole (Johnston 2004:15).

Despite mainstream criticism, government imposed regulations were initially the only viable option. Without requisite supporting institutions and after a century of unregulated rights to pollute, market-based solutions would have had a devastating impact on industry (Johnston 2004:7).

Government-imposed policies are advantageous if the environmental danger is imminent and widely understood. A clearly articulated standard can achieve a definite result. In addition, such policies enable all voices to be heard, including those without the ability to pay and of future generations.

A cogent and frequently circulated criticism of government-imposed policies is that they are inflexible and do not incorporate cost differentials among firms. Two facilities, for example, within the same industry and ostensibly similar, might be quite different from an engineering or technical perspective (Johnston 2004:17). Market-based solutions, such as pollution permits, enable low-cost firms to compensate high-cost firms, resulting in overall cost reduction.

Government-imposed policies are also criticized for not providing incentives to reduce emissions below the specified amount (assuming of course, that the firm is unwilling to inculcate such values on its own). In addition, government-imposed policies are accused of decreasing profitability and retarding technological development,

although little supporting evidence exists (Roodman 1998:27). The Montreal Protocol, for example, successfully forced private technological development, resulting in even faster timetables and stricter controls than government mandates (Anderson & Sarma 2002:362,265).

Taxes and Subsidies

Environmental taxes provide pecuniary incentives to nudge society toward desirable goals such as energy conservation, while discouraging harmful behavior such as environmental destruction. By paying a tax equal to inflicted environmental damage, polluters pay for the consequences of their actions and earns the right to pollute (Johnston 2004:8). Any revenue raised could offset income taxes, often cited as a hindrance to economic growth, which in turn could increase the palatability of an energy tax. Environmental taxes can also encourage production of substitutes. The total amount of emissions depends on the per unit fee: a higher fee, based on the law of demand, results in greater emissions reduction.

One example is the carbon tax, a per unit tax on carbon-based fuels, designed to reduce consumption of carbon-based fuels, while concomitantly generating less pollution and increasing the cost-competitiveness of non-carbon-based fuels.

Mainstream economics assumes that environmental damage can be internalized by twitching the right price signals, and once corrected, market participants will readily incorporate all requisite information. While such a position “may have been acceptable [when] economic activity only had a marginal influence on the environment” (Vant 2005:426), it is not acceptable today given the imminence of global warming and the complexity of environmental interactions (Lovelock 2000).

Green economics, while rejecting the growth-centered thrust of neoclassical economics, encourages a carbon tax as “the most powerful way of discouraging the use of carbon-based fuels” (Cato 1999: 81). But the tax is advocated as part of a holistic economic realignment emphasizing fairness, justice and environmental sustainability (Robertson 1999; Douthwaite 1992).

Environmental taxes are criticized because they increase prices and alter consumer behavior, distorting an otherwise equilibrium situation. Such a criticism however, is only valid if the pre-tax regime is pareto optimal; if not, taxes can nudge society toward more desirable goals. A second criticism is that such taxes are regressive since the poor spend a greater percentage of their income on energy and energy-intensive products (Roodman 1997:31). Third, since the optimal level of pollution is unknown, it cannot be ascertained a priori if a stated tax is too high or too low. Frequent adjustment of the tax rate could reduce business investment. Fourth, it is difficult to predict how consumers will react to a tax, rendering it problematic to achieve a stated outcome. Finally, and paradoxically, the more successful the tax, the less revenue generated.

A subsidy is a government payment, or tax break, designed to reduce a firm’s cost of production, which is otherwise too high to stimulate demand. A subsidy can either entice a firm to reduce production of an externality, or encourage production of an incipient technology that could ameliorate the externality.

Given a choice, mainstream economists prefer to tax than subsidize a negative externality, although theoretically both can achieve the same objective. A subsidy shifts an industry supply curve to the right, resulting in a greater amount of firms and higher industry output, while a Pigouvian tax shifts a supply curve to the left with a consequent

industry contraction (Cropper & Oates 1992:681). A tax reduces profits and consumption, whereas a subsidy reduces costs and increases profits.

Given a positive externality, subsidization can encourage achievement of societal goals by reducing costs and increasing competitiveness vis-a-vis competitors. A subsidy to wind producers, for example, could reduce costs, spur innovation and technological diffusion and lower prices.

Pollution Permits

A pollution permit is a licence or allowance to pollute issued by the state. The permits, which are de facto property rights, are either distributed free by the state or auctioned to businesses. The state sets an overall pollutant target lower than that of the unregulated market. Whereas with taxes, the state sets the price and the market determines the quantity of pollutants, with pollution permits the state caps the amount of pollution and the market sets the price.

The main advantage of pollution permits is the induced incentives: if a firm can reduce emissions below a specified level, it can earn credits for either future use or for sale, thus creating an incentive to reduce the pollutant more than the specified amount. Theoretically, a polluter with a high cost of pollution abatement will purchase a permit from a firm with a lower cost, although research to date fails to substantiate this claim (Fowline 2006:3).

Given a well-functioning market, polluters will trade permits until the price equals the marginal abatement cost, which should then equalize across all firms.

A second advantage is that the market itself is utilized to correct an externality by transmuting the social goal of environmental protection from a cost to a profit opportunity (Roodman 1998:6). It transmutes “a primal human impulse – greed – and redirects it

toward saving the planet rather than destroying it (Goodell 2006:36). Third, pollution permits economize on scarce information by shifting decision making from the state to the firm. Instead of the state gathering cost data on abatement sources, each firm decides whether to buy a permit, invest in pollution control equipment or buy the rights to pollute (Cropper & Oates 1992:686). Finally, permits motivate a firm to invest in additional pollution control technology since it can keep the resulting profit.

Pollution permits are enthusiastically endorsed by mainstream economists as the preferred method in fighting global warming. Samuelson and Nordhaus enthusiastically tout permits in the preface to their 25th edition of *Principles of Economics*, as “among the important innovations we survey” (2005:xviii).

Nevertheless, the state plays a crucial role. First, by creating the requisite market-supporting institutions; second by determining the ‘proper’ level of pollution; and third by defining property rights, thereby determining who has ‘the right to pollute.’

The United States pioneered pollution permits with the 1977 Amendments to the Clean Air Act, as a pro-market alternative to state-imposed regulations. The 1990 Amendments to the Clean Air Act established the sulphur dioxide reduction program, which is considered by many the archetypical success.

The Kyoto Treaty, which took effect February 2005 incorporated pollution permits as did a European wide emissions program also adopted in 2005.

Several exchanges have been established to facilitate the buying and selling of greenhouse gas emissions. The Chicago Climate Exchange, established in 2003 is a private, voluntary organization with over 175 members (Chicago Climate Exchange 2006).

And the European Union Emission Trading Scheme, established in 2005, is government sanctioned to prepare its members for the Kyoto Treaty (European Union 2006). Many economists are predicting that a global market in greenhouse gas emissions could become the largest commodities market in the world (Goodell 2006:30).

Despite the widely touted advantages of pollution permits, several criticisms exist. First, while markets can efficiently allocate goods based on preferences within an existing institutional context, markets cannot communicate deeply held societal values; this must be done by communication within the political process. Indeed, over-reliance on market forces can threaten the legitimacy of the governing process (Bulduc 2004:190). Reliance on market solutions ignores the strong ethical content of most environmental decisions which should be endogenous to the political process.

Second, a pollution permit regime emphasizes the narrow concept of efficiency, defined as the price equal to the incremental cost of producing the good, while ignoring ecological and social criteria, equally critical to societal welfare (Bulduc 2004:190).

Third, the exercise of a property right excludes two constituencies: those without the ability to pay and future generations. Since neither has a voice to affect present policy they are “powerless and vulnerable” (Streeten 1998:256). Without incorporating their preferences, internalization of the externality is incomplete, resulting in an inefficient allocation of resources (Dietz & Van der Straaten 1992:32).

Justifying this exclusion is the following narrow conception of a public good: ‘something common to a group of people which no one can be excluded from consuming.’ If, however, a more holistic definition of a public good is adopted such as ‘a public good is something to be enjoyed by

the entire community, present and future, able to pay and unable to pay” (Vatn 2005:420) then a legitimate claim for future generations can be made.

Mainstream economics discounts future preferences, in effect parrying an essential, ethical question: Do we have an obligation to future generations? Mainstream economics answers yes if substantiated by present consumer preferences. Otherwise, by maximizing current income, mainstream economics argue that future generations will inherit the pecuniary and technological means to ameliorate any problems. But what if today’s pollution decreases the regenerating capacity of the planet? Should not future generations be allocated a voice in present environmental policy to prevent this?

Fourth, it is not enough to assume superiority of private ownership since that reveals very little, “one must also list the more specific content of the rights and duties involved of the owner, non-owner and the state (Vant 2005:255). Pollution permits represent the latest commodification of resources, heretofore considered open and available to everyone. Nevertheless, the state has a role in injecting deeply held values that are ignored by private enterprise (Vant 2005: 260). Indeed, “markets may be wonderfully efficient systems, but they are no substitute for strong government action – both in setting the broad social goals of how to deal with global warming and, in the case of carbon markets, ensuring that the rules are not inclined in favor of private interests” (Goodell 2006:39).

Fifth, a pollution permits regime assumes that all costs are measured and that the price mechanism accurately conveys scarcity values. But many costs are unknown and even if known cannot be expressed in monetary terms.

Sixth, a pollution permit regime assumes that utility-maximizing behavior is

appropriate in all institutional contexts. But appropriate behavior in one institutional context such as the market, might be inappropriate in another context such as the environment. Mainstream economics is handicapped by advocating the use of microeconomic tools appropriate to the study of the firm for the study of the environment which requires a holistic worldview (Anderson 2006:20).

Seventh, mainstream economics disparages the subsidization of inefficient firms, yet under a pollution permit regime, technologically inefficient firms are allowed to produce via subsidies from more efficient firms. Should not future generations have a right to demand that such firms not be allowed to produce?

Eighth, although studies indicate that although overall reduction goals have been met, spatial inequalities have increased. In the United States, for example, total sulphur dioxide emissions fell from 1990 to 2001, but sixteen states saw an increase in emissions (Bulduc 2004). If pollution damage varies significantly within a region, an exposure-based, rather than an emissions-based program is appropriate. The former recognizes that damage from an additional pollution unit varies significantly across a region, so more permits are issued per unit of pollution; whereas the latter allows a firm to emit a certain quantity of pollutants regardless of the spatial source (Fowline 2006:3).

Ninth, pollution permits protect and enhance the status quo, which currently favors large fossil fuel corporations. This is a fixed cost barrier, preventing and discouraging potential firms from entering the industry. Supporting the status quo enables the dominant companies to influence the decision making process including when and if to introduce new technology.

Finally, the concept of a pollution permit is based on a disconnect between inputs, technology and the environment. Firms are allowed to produce an externality if they rectify it ex post, rather than adopt the requisite technology to prevent the externality ex ante? Pollution permits can only moderate environmental problems rather than ameliorate their rudimentary cause since the underlying cause market expansion and economic growth and economic growth independent of environmental sustainability (Wall 2006:210).

Conclusion

An important question for the 21st century is how far to extend the hegemony of markets. Are there some problems better solved by the state? How can markets account for voiceless consistencies?

Despite partisan rhetoric, the demarcation between market-based and state-based environmental policies is not so clear-cut, for each is “nothing more than man-made rules” (Swaney 1992:627). Both require articulation of society’s values to decide whose interests will be recognized and protected; and both are subject to manipulation by vested interests, “the latter however hides obscure vested interest more so than the former” (Dietz and Van der Straaten 1992:43).

A consensus has arisen that an efficacious solution to global warming requires the complementarity of markets and government, in order to use the strengths of each to compensate for the weakness of the other (Roodman 1997:24,27; ACCA 2003:30; Cropper and Oates 1992:700).

Can any lessons be drawn from the Montreal Protocol, considered “one of the world’s greatest achievements?” (Anderson and Sarma 2002:xxiii). Yes, one lesson is the need for an unprecedented level of cooperation, requiring short-term sacrifice

and forward thinking among the industrialized nations.

Second, it is necessary to develop an effective environmental consciousness. With the Montreal Protocol, “the mass media played a major role in shaping awareness and perceptions of the scientific, diplomatic, commercial and technical challenges of protecting the ozone layer . . . placing the ozone-layer issue directly in the public debate and on the policy-making agenda” (Anderson & Sarma 2002:290-91).

Raising environmental consciousness may also lower implementation and enforcement costs (Santopietro 1995) and allow a variety of solutions to be discussed. Pertaining to global warming, at least as of this writing, a sense of urgency is lacking. While there are many explanations, one is our own complacency, since for many, “global warming creates an illusion of a comfortable, warm future that is deeply appealing” (Flannery 2005:237).

We need a new conceptualization of globalization, one that fosters open communication, cooperation and community creation in order to develop institutions not seen before that match the problems we face (Vandt 2005:434). We also need a new credo allowing future generations input over present environmental policy, perhaps with an elected ombudsman and recognition that the present generation has a duty to make restitution for the harm we have already caused (Chong 2006:108,116). The objective is not to place pecuniary values on individual parts of the environment but to recognize these different pieces as a functional whole (Anderson 2006:21).

Selected References

Anderson, Steven and K. Madhava Sarma. (2002) *Protecting the Ozone Layer*. London: Earthscan Publications.

- Anderson, Victor. (2006) "Turning Ecobnomics Inside Out" *International Journal of Green Economics*, 1, 1/2, 11-22
- ACCA. (Association of Chartered Certified Accountants) (2003) *Environmental Taxes*. London: ACCA.
- Buldoc, Stephen. (2004) "Ceremonial Dimensions of Market-based Pollution Control Instruments: The Clean Air Act and the Cap and Trade Model", *Utilities Policy*, 12, 181-191.
- Cato, Molly Scott. (1999) "The Role of Ecotaxes in a Green Economy", in Molly Scott Cato and Miriam Kennett (Editors), *Green Economics: Beyond Supply and Demand to Meeting People's Needs*. Aberystwyth, UK: Green Audit Books, 78-86.
- Cato, Molly Scott and Miriam Kennett. (1999) "An Introduction to Green Economics", in Molly Scott Cato and Miriam Kennett (Editors), *Green Economics: Beyond Supply and Demand to Meeting People's Needs*. Aberystwyth, UK: Green Audit Books, 1-13.
- Chicago Climate Exchange. (2006) www.chicagoclimateexchange.com
- Chong, Chit. (2006) "Restoring the Rights of Future Generations", *International Journal of Green Economics*, 1 1/2, 103-120.
- Clinch, J.P. and M. Gooch. (2006) *Economic Instruments in Public Policy*. www.economicinstruments.com
- Coase, Ronald. (1960) "The Problem of Social Cost", *Journal of Law and Economics*, 3, 1-44.
- Demsetz, H. (1967) "Toward a Theory of Property Rights", *American Economic Review*, 57, 347-59.
- Dietz, Frank and Jan Van der Straaten. (1992) "Rethinking Environmental Economics: Missing Links Between Economic Theory and Environmental Policy", *Journal of Economic Issues*, 26, 1, 27-51.
- Douthwaite, Richard. (1992) *The Growth Illusion: How Growth has Enriched the Few, Impoverished the Many and Endangered the Planet*. Bideford, England: Green Books.
- Dragun, A.K. (1985) "Property Rights and Pigovian Taxes", *Journal of Economic Issues*, 19, 1, 111-122.
- European Union Emission Trading Scheme. (2006) ec.europa.eu/environment/climate/emission/
- Flannery, Tim. (2005) *The Weather Makers*. New York: Atlantic Monthly.
- Goodell, Jeff. (2006) "Capital Pollution Solution?", *New York Times Magazine*, July 30, 34-39.
- Hahn, R.W. and G.L. Hester. (1989) "Marketable Permits: Lessons for Theory and Practice", *Ecology Law Quarterly*, 16, 361-406.
- Johnston, Jason. (2004) "Tradeable Pollution Permits and Regulatory Game", April. www.law.uchicago.edu/Lawecon/workshop-papers/johnston.pdf.
- Klasassen, G. And A. Nentjes. (1997) "Creating Markets for Air Pollution Control in Europe and the USA", *Environmental and Resource Economics*, 10, 125-46.
- Krugman, Paul and Robin Wells. (2005) *Microeconomics*. New York: Worth.
- Kuttner, Robert. (1997) *Everything for Sale: The Virtues and Limits of Markets*. New York: Knopf.
- Lovelock, James. *Gaia: A New Look At Life on Earth*. Oxford: Oxford University Press.
- Mandelbaum, Michael. (2003) *The Ideas That Conquered the World: Peace, Democracy and Free Markets in the 21st Century*. New York: Public Affairs.
- McNeil, J.R. (2000) *An Environmental History of the Twentieth -Century World*. New York: W.W. Norton.
- Nash, J.R. and R.L. Revesz. (2001) "Markets and Geography: Designing Marketable Permit Schemes to Control Local and

- Regional Pollutants”, *Ecology Law Quarterly*, 28, 559-661.
- Pigou, Arthur. (1920) *The Economics of Welfare*. London: Macmillan.
- Pointing, Clive. (1991) *A Green History of the Earth*. New York: Penguin.
- Robertson, James. (1999) “A Green Taxation and Benefits System”, in Molly Scott Cato and Miriam Kennett (Editors), *Green Economics: Beyond Supply and Demand to Meeting People’s Needs*. Aberystwyth, Great Britain: Green Audit Books, 65-77.
- Roodman, David Malin. (1997) *Getting the Signals Right: Tax Reform to Protect the Environment and the Economy*. Worldwatch Paper 134. Washington, DC: Worldwatch Institute, May.
- Samuelson, Paul and William Nordhaus. (2005) *Economics*. 25th edition. New York: McGraw Hill.
- Santopietro, George. (1995) “Raising Environmental Consciousness versus Creating Economic Incentives as Alternative Policies for Environmental Protection”, *Journal of Economic Issues*, 29, 2, 517-524.
- Schmalensee, R. P.L. Joskow, A.D. Ellerman, J.P. Montero and E.M. Bailey. (1998) “An Interim Evaluation of Sulfur Dioxide Emissions Trading”, *Journal of Economic Perspectives*, 12, 53-68
- Singer, Joseph William. (1998) “Property”, in David Kairys (Editor), *The Politics of Law: A Progressive Critique*. Third Edition. New York: Basic Books, 1998, 240-258.
- Streeten, Paul. (1998) “What Do We Owe the Future?”, in Charles Wilber (Editor), *Economics, Ethics, and Public Policy*. Lanham, Maryland: Roman and Littlefield, 251-266.
- Swaney, James. (1992) “Market versus Command and Control Environmental Policies”, *Journal of Economic Issues*, 26, 2, 623 – 633.
- Tietenberg, T.H. (1998) “Ethical Influences on the Evolution of the US Tradeable Approach to Air Pollution Control”, *Ecological Economics*, 24, 241-57.
- Torres, Gerald. (1998) “Environmental Law”, in David Kairys (Editor), *The Politics of Law: A Progressive Critique*. Third Edition. New York: Basic Books, 1998, 172-189..
- Vatn, Arrild. (1997) “Externalities: A Market Model Failure”, *Journal of Environmental and Resource Economics*, 9, 135-51.
- Vatn, Arild. (2005) *Institutions and the Environment*. Cheltenham, UK: Edward Elgar.
- Wall, Derek. (2006) “Green Economics: An Economics and Research Agenda.” *International Journal of Green Economics*, 1, 1/2, 201-14.

Jack Reardon
School of Business
Hamline University
St. Paul,
Minnesota, USA
jreardon02@gw.hamline.edu.

Principal-Agent Theory

Andreas Feidakis

Introduction

According to classical finance theory, corporations are formed mainly by two separated parties: the shareholders and the managers. The shareholders (the principal) employ and entrust tasks to professional managers (the agent), since they neither have the time nor the ability to do the task themselves, and thereby they separate the corporation owners from control of the business. This principal-agent relationship implies the existence of conflicts of interest between these two parties. When the top executives are not the owners of the company, it cannot be expected that they will watch over it with the same anxious vigilance as the owners themselves (Smith 1776). Pondy (1969) finds that firms which are controlled by managers are characterized by higher levels of administrative complexity than those controlled by the owners.

Berle and Means (1932) report that the ownership of large companies is fragmented and dispersed among many shareholders, who are lacking the power to constrain managerial discretion. Professional managers with little or no equity gained the management responsibility for firms that had previously been managed by their owners. The separation of roles between ownership and control raised the possibility that both parties may have divergent interests. Chandler (1977) finds that the more complex and larger the firms are in US the more their managers are separated from their ownership group.

Agency problems are said to occur when agents pursue individual goals, which are not necessarily consistent with those of the company. Shareholders undoubtedly want managers to pursue policies that maximize their wealth, but managers may have other

personal objectives (Marris 1963; Williamson 1975). Managers may pursue growth (Grabowski & Mueller 1972; Cubbin & Leech 1986), perquisite seeking (Williamson 1963; Jensen & Meckling 1976), personal power and prestige (Simon 1945), different time horizons (Smith & Watts 1983) and risk reduction (Amihud & Lev 1981; Jensen 1986) instead of maximizing shareholder wealth.

Agency costs are incurred when shareholders have to spend time and money resolving these conflicts. Agency costs include not only the monitoring and controlling costs but also the costs of providing managers with incentives to maximize shareholders wealth. However, agency problems can never be perfectly solved and so shareholders experience residual losses due to divergent behavior of the managers (Jensen & Meckling 1976). Minimization of agency costs means maximization of the value of the firm.

Jensen and Meckling (1976) define the principal-agent relationship as a contract under which the principal engage the agent to perform some service on its behalf which involves some decision making authority to the agent. Thus, the principal-agent relationship is a contracting problem concerning how much of the value that the agent produces should go back to him in the form of a payment.

Moreover, risk preferences of agents are different from those of the principal, resulting in less than optimal decision making. Owners pursue corporate goals and are risk neutral because they are likely to be diversified across the securities of many firms (Fama 1980), whereas agents pursue personal goals and are risk averse. Managers will make decisions that minimize risk in order to assure steady income and continued employment (Donaldson & Lorsch 1983; Jones & Butler 1992; Eisenhardt 1989; Stiglitz 1987).

Another aspect of the principal-agent problem is the additional assumption of asymmetric information, meaning that the agent knows more than the principal about the business. When there is asymmetric information then the principal-agent model implies two problems: moral hazard and adverse selection.

A moral hazard problem appears when the principal has difficulty in observing the actions and the effort of the agent. This situation exists after the making of the contract between principal and agent. A moral hazard problem emerges because the shareholders cannot certainly observe in detail whether the management is making the appropriate decisions (Arrow 1985). Moral hazard refers to the problem of inducing agents to supply proper amounts of productive inputs when their actions cannot be observed and contracted directly (Holmstrom 1982).

An adverse selection problem appears when the principal has difficulty in discovering the true nature of the agent. This situation exists before the making of the contract between principal and agent. An adverse selection problem emerges because the principal cannot check whether his agent actually complies with his contract or the agent misrepresents his abilities. Adverse selection refers to a situation where actions can be observed, but it cannot be verified whether the action was the correct one, given the agent's contingency, which he privately observes (Holmstrom 1982).

In general, according to adverse selection the principal must rely on proxy indicators of success rather than success itself (adverse selection), while according to moral hazard the agent directs attention toward the satisfaction of proxy measures rather than toward the success of the task itself (Bromley 1989).

Distribution of information from big listed firms to their shareholders takes place through annual reports and announcements through the press. The content of these information sources is regulated by legal disclosure requirements. The aim is to reduce problems arising from asymmetric information and insider trading. Several EU-directives contain provisions concerning the disclosure of accounting information. The separation of management and ownership is, of course, more characteristic in big listed firms than in smaller ones.

There is a particular field in the literature of Finance that focuses on the principal-agent relationship, corporate governance. Corporate governance investigates how to protect the interests of shareholders and how to secure and motivate efficient management of corporations by the use of incentive mechanisms, such as contracts, organizational designs and legislation (Shleifer & Vishny 1997; Hart 1995; Williamson 1984). Corporate governance can be expected to reduce agency problems.

Control of Managerial Behavior And Optimal Ownership Structure

Sappington (1991) highlights that the central concern in principal-agent relationship is how the principal can best motivate the agent to perform as the principal would prefer, taking into account the difficulties in monitoring the agent's activities.

Shareholders use several control mechanisms in order to be in command of the managerial actions. They need a control system that provides measures of performance and specifies the relationship between rewards and punishments and the measures of performance. The shareholders delegate to the board of directors the task of protecting the interests of shareholders, by selecting, hiring, controlling, firing,

compensating and disciplining the management team.

The more independent from top management are the directors (e.g. outside directors), the more efficient is the agent monitoring (Fama 1980). Independent directors implement hostile takeover defenses to increase the bargaining position of target shareholders while management-dominated directors tend to implement defenses that increase the entrenchment (Cotter et al 1997). Also, outside boards are more likely to remove CEO's as a result of poor company performance (Byrd & Hickman 1992).

In particular, the bigger the percentage of a company's stock owned by institutional investors, the better the monitoring of the performance of the managers and directors. Hill and Snell (1989) find a positive relationship between ownership concentration and productivity. Glassman and Rhoades (1980) find a positive relationship between ownership concentration and profit. Mehran (1995) provides evidence of a positive relationship between managerial equity ownership and firm performance in line with Wruck (1988). Hartzell and Starks (2003) find a significantly negative relationship between the concentration of institutional ownership and the level of management compensation.

Jensen and Meckling (1976) state that the value of the firm increases together with the ownership concentration on condition that the costs incurred by increasing ownership concentration are smaller than the benefits derived from reducing agency costs. Fama and Jensen (1983) claim that if the concentration of ownership of the firm rises such that it allows entrenchment of the management then the value of the firm falls. According to this point of view, the management might be tempted to adopt not the optimum decision making as it will not be

exposed to the disciplining role of the market for corporate control.

On the other hand, Shleifer and Vishny (1986) state that for diffusely held firms it might not be in the interest of any of the shareholders to monitor management (free-rider problem) in accordance with Grossman and Hart (1980) and Leland and Pyle (1977). The presence of a large minority shareholder makes the occurrence of a takeover more likely and therefore increases the value of the firm.

Large shareholders also have the power to implement management changes and facilitate takeovers (Grossman & Hart 1980; Shleifer & Vishny 1986). The dilemma for a large shareholder is the optimal risk diversification under a fully dispersed ownership or the optimal monitoring incentives under a concentrated ownership. However, in liquid secondary markets large shareholders would rather sell their stake in mismanaged firms than try to fix the management problem (Mayer 1988; Black 1990). Mudambi and Nicosia (1998) find that ownership concentration and the degree of shareholder control have different effects on firm performance. Increased shareholder control appears to be positively related to performance but increased concentration seems to be inversely related to the same performance measure. The same authors find a non-monotonic relationship between managerial share holding and firm performance. In contrast, Demsetz and Lehn (1985) show that firms where ownership is widely distributed have profits as large as those with concentrated ownership in line with Demsetz (1983) and similar to Tsetsekos and DeFusco (1990). Demsetz and Lehn (1985) find that the financial performance of a firm does not depend on its ownership and its corporate control. Dalton et al (2003) find that there is no meaningful relationship

between the extensiveness of insider ownership and firm performance.

They give incentives to the managers to pursue the shareholders' goals with contracts and arrangements for compensation such as promotions (Lazear & Rosen, 1981; Sappington 1991) bonuses, stock options plans (Beck & Zorn 1982; Bhagat et al 1985; Haugen & Senbet 1981) and performance stocks. By giving to managers, stock options or stocks, their interest will become more closely aligned with the stockholders; it may also induce them to take on more risk (DeFusco et al 1990; Sappington 1991; Sanders 2001). Masson (1971) declares that firms with compensation packages that emphasize stock market performance outperform firms without such an arrangement. Moreover, Murphy (1985) reveals that salary, bonus and total compensation are positively related to shareholder return and growth in sales in line with Abowd (1990). The salary component of managers' compensation can provide insurance against market forces beyond their control (Sappington 1991). However, there is a difficulty in observing the connection between managerial behavior and stock options (Williamson 1970). What is more, some researchers argue that, while financial incentive schemes improve productivity in principle, in practice they induce significant side effects that are costly to employee morale and productivity (Hamner 1975; Beer et al 1984). Besides, while managers can seek higher income by increased effort, according to the pay-for-performance system, it is often easier and less demanding to influence their outcome even by distorting and falsifying the figures (Becht et al 2002).

The fear of a hostile takeover puts pressure on the management to take actions that will maximize stock prices (Fama 1980; Williamson 1970; Grossman & Hart 1980). Takeovers are useful both because they

reduce the informational monopoly of the managers and because they allow for the replacement of inefficient managers by replacing the owners (Manne 1965). Gompers et al (2003) find that the more anti-takeover provisions a firm has in its charter, the lower its performance is. Jensen (1988) asserts that efficient capital markets render the principal-agent issue largely irrelevant. Capital markets serve as a metering device on managerial performance. Jensen (1988) declares that takeovers occur because managers act in their own interests to the detriment of shareholders. The market will punish managers who do not act in the best interests of shareholders via takeovers. The threat of bankruptcy with its consequent damage to their reputation may also discipline managers.

Fama (1980) asserts that if managers act for their own benefit or act in other ways that fail to increase firm's value, their behavior will be reflected in depressed share prices. Therefore, the board of directors can reward or penalize them depending on their abilities or not to increase shareholders wealth. Stock prices serve as a convenient and inexpensive monitor of the management's performance. Direct monitoring and evaluation of the daily activities of an executive is problematic given the complex nature of activities an executive performs (Sappington 1991).

Rivalry in managerial labor markets helps the effectiveness of incentives and monitoring of managers (Fama 1980).

The level of shareholder protection at the country level, the degree of board independence and the existence and independence of board committees seem to play a more important role in firm performance than other corporate governance practices (Bruno & Claessens 2007).

Jensen (1986, 1989) argues that the best way to resolve the agency problem between the managers and shareholders is to have the firm take on as much debt as possible. This

would limit managerial discretion by minimizing the free cash flow available to managers and, thus, would provide protection to shareholders. The main difficulty with Jensen's control theory is that highly levered firms may face direct bankruptcy costs or indirect costs in the form of debt-overhang (Myers 1977; Hart and Moore 1995). To reduce the risk of financial distress the firm may rely partly on equity financing. In the bank-based systems, free cash flow investments are made in low-risk projects and are acceptable to the bank monitors since they reduce the probability of financial distress (Cohen & Boyd 2000).

Burkart et al (1997) argue in particular that too much monitoring and legal protection may hurt managerial incentive and consequently generate lower returns and valuation. The managers are less inclined to show initiative when shareholders are more likely to interfere. In the same way, Boot et al (2006) find that corporate governance controls may prevent management from doing what it wants, thus exacerbating agency problems. There is a trade-off between the gains from monitoring and those from managerial initiative and excessive monitoring can therefore be inefficient.

La Porta et al (1999) underline the importance of a legal approach to corporate governance. When the rights of minority shareholders are protected through laws the investors are willing to finance firms, encouraging the development of equity markets. Strong investor protection is positively associated with effective corporate governance. Agency problems can be mitigated by means of legal protection of minority investors, the use of board of directors as monitors of the senior management, also with an active market for corporate control (Gompers et al 2003).

Rediker and Seth (1995) highlight that there is a complex interaction of different

corporate governance methods that may substitute for each other. For example, a firm with high ownership concentration is not likely to be taken over and hence ownership concentration might be said to substitute for the threat of takeover as a governance device.

Last but not least, rational principals will only pursue the available techniques for control to the point that the marginal increment in agency costs rise to equal the marginal benefits to them of the additional increment in faithfulness that they produce. That is to say, sometimes it is cheaper for principals to endure a certain amount of dereliction of duty by their agents than paying for the precautions needed to prevent or punish it.

Implications

We can see that principal-agent relationship has many and important implications that play crucial role for the business of a firm.

One implication is the insider trading problem. Insider trading occurs when an employer of the firm (the insider) uses information about the firm that is not public. Insider trading is generally prohibited. Firms do not allow inside trading because shareholders are unable to control efficiently the activity of their employees as insiders (Easterbrook 1985, Ross 1979).

Lins and Warnock (2004) find that a firm's corporate governance environment, at both the firm and country level, is directly related to the willingness of a large and sophisticated group of foreign investors to hold its shares. Firms, whose managers have sufficiently high control rights that they may be expected to expropriate minority equity investors, attract significantly less investment, especially in countries with poor external governance.

La Porta et al (1997) and Giannetti and Koskinen (2004) argue that improved corporate governance should be associated

with improved equity market development as local outsiders increase participation and the international integration could increase when these new equity investors diversify internationally.

Kock et al (2005) find strong support that different corporate governance mechanisms can serve as important tools for public and corporate board level efforts to enhance environmental performance. Anti-takeover amendments and provisions that restrict managers' personal liability to create a sphere of bad discretion allow managers to shirk by under investing in potentially financially beneficial levels of environmental performance. On the other hand, corporate governance structures that emphasize higher levels of performance pay and lower degrees of monitoring create a degree of good discretion that enhances environmental firm performance.

Another implication is the management compensation. It is suggested that one way to reduce the agency conflicts is by compensation policy. General managers emphasize that the level of compensation is a very important symbol of status (Donaldson & Lorsch 1983; Kotter 1982). In the absence of managerial compensation, managers may be overly concerned with job security. On the contrary, pay-for-performance systems concentrate the attention of managers on compensation rather than on effort and no manager has any incentive to inform the shareholders if these goals are reached by illegitimate or even illegal means, since they all benefit from the system. The result of this arrangement is an explosion of managers' compensation (Frey 2003; Osterloh & Frey 2000; Murphy 1999). Stock options plans create major conflicts of interest when the CEO's borrow from the firm to purchase their stock options (Becht et al 2002). Besides, incentive pay might cause managers to focus excessively on short-run profits and ignore

the long-run value of the firm. Additionally, compensation can be based on the agent's performance relative to that of other agents. Such schemes, which use ordinal measures of performance rather than cardinal measures, are called tournaments (Mookherjee 1988). Promotion-based incentives are a commonly observed example of tournaments (Baiman 1990). Relative performance evaluation can be helpful in reducing moral hazard costs, because it provides better risk sharing (Holmstrom 1982). In practice, though, it is often difficult to measure the agent's performance precisely (Sappington, 1991). Opportunistic behavior by managers reduces the potential willingness of investors to finance the firms (Grossman & Hart 1986; Williamson 1985).

One more issue is related to the nature of teamwork. When agents are placed on individual pay-for-performance schemes, they may not like to help and cooperate with their coworkers (Drago & Garvey 1998). Agents have a strong incentive to collude against the principal, but no incentive to help each other (Nilakant & Rao 1994). Where output reflects the contribution of many individuals there is a problem in identifying and metering individual contributions (Alchian & Demsetz 1972; Nilakant & Rao 1994; Holmstrom 1982).

A further implication is risk taking. In order to minimize their risk exposure, managers may adopt detrimental entrenching methods, such as compromising performance measures, neutralizing control mechanisms, or adopting deleterious corporate strategies (Walsh & Seward 1990). Amihud and Lev (1981) claim that managers may use conglomerate mergers, which are often associated with negative shareholder returns, simply to reduce employment and earnings risk. DeFusco et al (1990) provide evidence that stock option plans have asymmetric payoffs that could induce managers to take on

more risk and may induce a wealth transfer from bondholders to stockholders.

An additional implication is the capital structure of the firm. Barnea et al (1981) argue that financing preferences may result from agency costs associated with debt. Myers and Majluf (1984) show that asymmetry of information between managers and outside investors may affect the decision making for the investment financing of a major investment-corporate acquisition.

Atkinson and Galaskiewicz (1988) say that manager controlled firms are more generous in terms of contributions to charity than owner controlled ones.

Corporate governance is more important to companies that rely heavily on external financing. This is because corporate governance acts as a signaling device of positive NPV projects, thus allowing a more efficient capital allocation. Once the funds have been allocated and the signaling role ends, corporate governance still ensures its positive role through the monitoring of the management (Bruno & Claessens 2007).

More importantly for principal-agent theory, the financial markets have been plagued by huge scandals relating to excessive manager compensations and fraudulent book keepings.

Principal-agent theory is part of the explanation of project escalation decisions in individualistic, utilitarian-involved Western cultures, but is at best a very weak explanatory theory in collectivist, loyalty-involved cultures. In collectivist cultures, society itself imposes significantly higher costs – such as censure or ostracization – on individuals who violate collective norms by acting in their own interests rather than of the in-group (Sharp & Salter 1997).

Lhuillery (2006) finds that firms with corporate governance practices that are shaped in order to defend shareholders' rights are more R&D intensive. The higher the

shareholders are taken into consideration by the managers, the higher the R&D investments will be.

Easterbrook (1984) suggests that dividends may be useful in reducing the agency costs of management. Dividends may keep firms in the capital market, where monitoring of managers is available at lower cost and may be useful in adjusting the level of risk taken by managers and the different classes of investors.

According to agency theory, in the absence of counteracting incentive compensation or shareholder pressure, managers will tend to take on less business risk than would be optimal for shareholders. Stock options can create a powerful incentive to increase risk taking because the payout structure provides a leveraged gain if the risk-taking strategy succeeds and no loss if the strategy fails (Murphy 1999).

Another important variable affecting the impact of principal-agent relationship is size. In the corporate governance literature, some theories and empirical evidence imply that strong corporate governance is more beneficial for larger firms than for small ones (Bruno & Claessens 2007). Excessive regulation can be especially counterproductive for small firms where the agency costs more likely outweigh the benefits (Holmstrom & Kaplan 2003; Chhaochharia & Grinstein 2005). Strong corporate governance can be costly, limit managerial freedom of initiative and thereby negatively affect firm's performance (Bruno & Claessens 2007).

Corporate governance has significant implications for the growth prospects of an economy. Proper corporate governance practices diminish risk for investors, attract investment capitals, improve corporate performance, affect investors' confidence and retain economic stability (World Bank 1998). What is more, Johnson et al (2000) and

Mitton (2001) report how weak corporate governance worsened the 1997 Asian currency crises.

However, in recent years, another form of principal-agent relationship has drawn attention, controlling shareholders taking actions that are for their own benefits at the expense of minority shareholders (La Porta et al 1998). Such expropriation of minority shareholders by controlling shareholders takes a variety of forms, such as dilutive share issues, transfer pricing between related firms, etc. Johnson et al (2000) use the term 'tunneling' to describe the transfer of resources out of firms for the benefits of their controlling shareholders. Much evidence emerging during the Asian crisis shows that tunneling is a serious problem in emerging markets. The recent debacles of Enron, Worldcom, and Global Crossing convince that tunneling is possible even in developed countries.

Principal-Agent Relationship around the World

There are two major corporate governance systems, the market-based and the bank-based system. There are major differences across these two systems, with U.K. and U.S. rules designed mainly to promote shareholders' value (market-based system) at one extreme, and German or Japanese rules designed to balance the interests of shareholders and employees (bank-based system) at the other. U.S. and U.K. follow the Anglo-American outsider system of control which is a securities market-based one. The market-centered system is characterized by a widely held nature of securities and the lack of significant intercorporate, individual, family and bank holdings. Germany follows the Continental European insider system of control which is a bank-based one. In Germany, the ownership of a firm can be distinguished from control. The separation of

ownership and control in the publicly owned corporations prevails in the U.S. and the U.K., in which no single shareholder owns more than a small fraction of a corporation's stock.

The market based system seems the most effective in monitoring and disciplining management. The Anglo-American system is based on the powers and obligations of the board of directors and the market intervention through takeovers. The market based system provides flexibility to firms to use new structures and financing forms and encourages the growth of firms based on knowledge as opposed to hard assets.

The bank-based system does not appear to provide appropriate monitoring. The bank-based system faces the lack of external control mechanisms such as a well functioning capital market, a competitive market for corporate control and a competitive market for managers. Thus, any solution to the principal-agent problem must rely on internal mechanisms (Cohen & Boyd 2000).

In Germany, the equity market is not well developed and hence there is no monitoring function. There are concentrated holdings through intercorporate investments or concentrated voting in the banks that make the market for corporate control not operational. The presence of blockholdings does not generally imply shareholder maximization (Shleifer & Vishny 1997). Large blockholders have incentives to maximize the value of their shares and this does not always involve maximization of the firm value on behalf of all shareholders, especially the small shareholders. Management compensation contracts based on firm's financial performance are generally not observed. Managers receive a fixed salary and so they prefer to take low-risk projects and reduce the risk of financial distress of the firms.

Any monitoring activity must come from the supervisory board, the intercorporate shareholders or the banks. The labour component of the board is interested in continuity of employment for employees and as a result it prefers low risk corporate activities. The non-labour component of the board is elected by shareholders and given that banks have the controlling vote, this group represents the banks' interests. Due to their information advantage, banks are potentially effective monitors, but generally have little incentive to act on behalf of other shareholders. The banks are interested in firm's liquidity and the stability of cash flows. The banks use their votes to maximize the value of debt and not that of equity. The banks use their voting rights in the interest of management as opposed to that of shareholders when banks have debt or free-related income associated with the firm. The final monitoring group is the intercorporate shareholders who regard their investment as a way of ensuring business and not as an actively managed portfolio. In all, the supervisory board may not be structured to provide the necessary monitoring (Cohen & Boyd 2000).

In Japan, the shareholdings of Japanese companies are divided among financial institutions, which include banks, trust banks, life companies, pension and mutual funds, intercorporate holdings and individual shareholdings. Managerial compensation contracts based on share price performance are rare in Japan and are inconsistent with cultural norms in which team and not individual effort is valued.

The board of directors is ineffective in its monitoring role. In US, many directors are appointed from outside of the company and they are required to have general expertise in corporate management in order to monitor the operation of the company on behalf of the shareholders. In Japan, the composition of the

board is different; the directors are usually promoted from the ranks of middle managers of the company and in many cases continue to be its employees. The potential for promotion to the board creates an important incentive for the employees to be loyal and work hard for the company. However, the board has a hierarchical structure which can undermine its supervision of the affairs of directors, especially the president. Second, the number of directors tends to increase over time in order to reward long-serving managers and the boards are too large to maintain flexibility. Thus, in spite of the legal responsibility, the role of the board of large Japanese companies tends to be superficial as regards not only supervising the affairs of directors, but also directing the management of the company (OECD 2001). In Japan the number of outside directors is simply small and is more or less coterminous with the company's senior management (Lincoln et al 1992).

The shareholders do not have any control and they do not monitor the management through the annual shareholders' meeting (Baum & Schaefer 1994). There is a lack of serious questioning of managerial decisions by the board. The large shareholders may generally refrain from intervening in the management of the company, because they put more importance on maintaining or even expanding their business with the company rather than seeking the short-term returns on shares (OECD 2001). The pension and mutual funds and life companies prefer to waive all or part of their rights as shareholders for the sake of stable shareholding agreements and mutual trust. The intercorporate shareholders are interested in maintaining reciprocal insurance in the event of financial problems and generating business ties. Any attempt to monitor a manager and agitate for change or dismissal, except in the extreme situation of financial

distress, would be seen as a violation of maintaining business ties. Furthermore, banks are concerned with the possibility of the financial distress of the firm and liquidity problems. They are interested in minimizing the default risk even if this intervention in this regard does not maximize their shares' returns. Finally, even the group that has the largest influence on the day-to-day operations of the firm, the 'old boy network', is not necessarily consistent with non-corporate shareholders' wealth. It concerns the impact of corporate decisions on workers or on the economy and not on the non-corporate shareholders (Cohen & Boyd 2000).

Hanazaki et al (2004) argue that even though banks and insurers are endowed with substantial power to vote against the bank management, they are less effective monitors for two reasons. First, they have business relations with the banks they hold shares in. Second, they themselves appear to have weak corporate governance. They conclude that banks and insurance companies collude with the bank management.

Unlike in the Anglo-American system, in practice the mechanism to discipline managers through hostile takeovers has not functioned well in Japan. This is not because of the legal framework, which is less restrictive than that of the US as it does not provide the incumbent management with various anti-takeover measures, but mainly because of the stable ownership structure supported by cross-shareholdings (OECD 2001). The stable cross shareholding is aimed to prevent hostile takeovers and to ensure that control is exercised by the shareholders rather than by third parties (Sheard 1996). In Japan the individual shareholders are more or less insignificant.

Equally crucial is the fact that, outside the U.S. and the U.K., the corporation with widely dispersed ownership is not the rule but the exception (La Porta et al 1999). What

prevails throughout the developing world is the corporation with concentrated ownership, dominant stockholders ("blockholders") who directly control managers. Only in economies with good shareholder protection there are firms widely held by many shareholders. In most countries, instead, firms are controlled by families or by the State.

Also widespread in the developing countries is the use by dominant shareholders of means to control corporate assets considerably greater, even than those to which their stock ownership rights would directly entitle them. According to this phenomenon, the effective control rights exceed nominal cash-flow rights.

The key potential conflict of interest in the developing world as a whole therefore tends not to be between managers and shareholders but between dominant owner-managers on one hand and minority shareholders and other investors (domestic and foreign) on the other. This conflict of interest is commonly referred to as the "expropriation problem", as opposed to the agency problem.

Criticism of Principal-Agent Theory

Several authors criticize the generalizability of the principal-agent model. Kaplan (1983) questions whether managers engage in continuous individual maximizing behavior. Nilakant and Rao (1994) argue that even in an economic sense principals can enhance their welfare by investing in coordination, consensus-building and legitimacy-seeking activities rather than by investing in monitoring mechanisms and compensation systems. There are researchers that support the view that pay is not very closely related to performance (Lawler 1971; Medoff & Abraham 1980). Many psychologists claim that monetary rewards are counterproductive (Deci 1972; Slater 1980; Kohn 1988). Performance outcome depends not only on the person executing the effort but also on

others on whom the person is interdependent. The higher the extent of facilitative effort, the weaker the link between compensation and individual compensation. Organizational performance is determined by both operational and facilitative efforts. Operational effort can be observe, monitor and hard to misrepresent. Facilitative effort is difficult to observe and hard to monitor. Given this, it is likely that agents will under-invest in facilitative effort and overemphasize operational effort. Other agents who must act like principals have little incentive to monitor facilitative effort. Managers have more incentives to give uniform and good performance evaluations to subordinates, minimizing the monitoring costs and in return for the cooperation of employees. In addition, managers have incentives to build networks of trusted supporters in order to execute the facilitative effort (Nilakant & Rao 1994). A better option is to reward employees on the basis of total output, outcome based compensation contracts like profit-sharing plans, which provides incentives for both operational and facilitative effort (Baker et al 1988).

Simon (1991) argues that the moral hazard and opportunism assumptions have led agency theory to incorrectly rely on monitoring and compensation as the only remedies for self interested behavior. Authority, loyalty, identification with organizational goals and good spirited coordination are also important for motivating the effort required for firm success. Compensation is only one element of a set of valued returns that motivate an agent's actions. Moreover, the principal-agent model exaggerates the aversion of the agent to work. It is more useful to view a company as a network of interdependent roles rather than as a nexus of contracts. In a world of interdependencies, facilitative effort is more critical than operational effort. It is very

difficult for the principal-agent model to motivate, monitor and reward facilitative effort. Perhaps the designing structures of work around collaboration and trust (Orsburn et al 1990; Trice & Beyer 1984) are better than designing contracts based on suspicion and mistrust.

Principal-Agent Theory And Public Governance

Public governance uses a completely different system to align agents' behavior to principals' goals. It relies on self-selection and socialization to bring about an internalized intrinsic behavior of agents consistent with the goals of the public organization. The approach on which public governance is based places more emphasis on guiding agents' behavior by intrinsic motivation, while agency theory seeks to produce this alignment exclusively by setting the right extrinsic (monetary) incentives (Frey 2003).

Moreover, public governance also relies on external incentives to align agents' behavior with the goals of the organization. It gives titles to agents to clearly indicate their place in their hierarchy. It provides to the agents prestige and visibility in the public administration, which is an important extrinsic reward.

On the other hand, the absence of a market for shares removes sanctions for bad performance and incentives to managers in the public sector. Civil servants, who are not entitled to the gains due to improved efficiency, have no financial incentive to monitor public firms (OECD 1998). Public firms sometimes have unlimited access to government funds, so the firm runs no risk of bankruptcy, no matter how inefficient is its management (Kornai 1980). Management may also take advantage of deficient monitoring to shirk its duties and allows the establishment of high wages and overemployment (Pryke 1981).

Summary and Conclusions

Principal-agent theory not only dominates the academic literature, but has been accompanied by major applications in business practice. The principal-agent relationship has implications in the most important matters of corporate finance like firm's performance, risk, management compensation, corporate governance, dividend policy, capital structure, merger and takeovers and so on.

Corporate governance tries to solve the principal-agent problem and exerts a positive effect on firm's performance. This is because corporate governance acts as a monitoring and discipline device, ensuring that management pursues shareholders' goals. Principal-agent theory affects the real economy and an optimal corporate governance model accelerates economic growth.

Major corporate failures at Enron, Vivendi, Arthur Andersen, Marconi, WorldCom and others have put corporate governance in the headlines. They remind us of the catastrophic consequences of poor corporate governance even in the most highly developed countries.

The bottom line, however, is that the principal-agent problem can never be 100% solved in a world where virtually everyone looks after his own self-interest and the relevant information for evaluating performance is imperfect, costly to obtain and unequally distributed between the agent and the principal.

Selected References

Abowd, J.M. (1990) "Does Performance-Based Managerial Compensation Affect Corporate Performance?", *Industrial and Labor Relations Review*, Volume 43, pp. 528-738.

- Agrawal, A. and G.N. Mandelker. (1987) "Managerial Incentives and Corporate Investment and Financing Decisions", *The Journal of Finance*, Volume 42, Number 4, pp. 823-837.
- Alchian, A.A. and H. Demsetz. (1972) "Production, Information Costs, and Economic Organization", *American Economic Review*, Volume 62, Number 5, pp. 777-795.
- Amihud, Y. and B. Lev. (1981) "Risk Reduction as a Managerial Motive for Conglomerate Mergers", *Bell Journal of Economics*, Volume 20, pp. 605-617.
- Arrow, K.J. (1985) "The Economics of Agency" in John W. Pratt and Richard J. Zeckhauser (Editors), *Principals and Agents: The Structure of Business*. Boston, Mass: Harvard Business School Press, pp. 37-51.
- Atkinson, L. and S. Galaskiewicz. (1988) "Stock Ownership and Company Contributions to Charity", *Administrative Science Quarterly*, Volume 33, pp. 605-617.
- Baiman, S. (1990) "Agency Research in Managerial Accounting: A Second Look", *Accounting, Organizations and Society*, Volume 15, Number 4, pp. 341-371.
- Baker, G.P., M.C. Jensen and K.J. Murphy. (1988) "Compensation and Incentives", *The Journal of Finance*, Volume 43, Number 3, pp. 593-616.
- Barnea, A., R. Haugen and L. Senbet. (1981) "An Equilibrium Analysis of Debt Financing Under Costly Tax Arbitrage and Agency Problems", *The Journal of Finance*, Volume 36, pp. 569-581.
- Baum, H. and U. Schaede. (1994) *Institutional Investors and Corporate Governance in Japan*. Berlin: Walter De Gruyter.
- Becht, M., P. Bolton and A. Röell. (2002) *Corporate Governance and Corporate Control*. ECGI Working Paper Series in

- Finance 02/2002, 30 September 2002. Brussels: European Corporate Governance Institute.
- Beck, P.J. and T.S. Zorn. (1982) "Managerial Incentives in a Stock Market Economy", *The Journal of Finance*, Volume 37, pp. 1151-1167.
- Beer, M., B. Spector, P.R. Lawrence, D.Q. Mills, and R.E. Walton. (1984) *Managing Human Assets*. New York: The Free Press.
- Berle, A. and G. Means. (1932) *The Modern Corporation and Private Property*. New York: Macmillan.
- Bhagat, S., J.A. Brickley and R.C. Lease. (1985) "Incentive Effects of Stock Purchase Plans", *Journal of Financial Economics*, Volume 14, pp. 195-215.
- Black, B.S. (1990) "Shareholder Passivity Reexamined", *Michigan Law Review*, Volume 89, pp. 520-608.
- Boot, A.W., R. Gopalan and A.V. Thakor. (2006) "The Entrepreneur's Choice Between Private and Public Ownership", *The Journal of Finance*, Volume 61, pp. 803-836.
- Bromley, D.W. (1989) *Economic Interests and Institutions: The Conceptual Foundations of Public Policy*. Oxford: Basil Blackwell.
- Bruno, V.G. and S. Claessens. (2007) *Corporate Governance and Regulation: Can There Be Too Much of a Good Thing?*, World Bank Policy Research Working Paper 4140. Washington DC: World Bank.
- Burkart, M., D. Gromb and F. Panunzi. (1997) "Large Shareholders, Monitoring and The Value of The Firm", *Quarterly Journal of Economics*, Volume 112, pp. 693-728.
- Byrd, J.W. and K.A. Hickman. (1992) "Do Outside Directors Monitor Managers? Evidence From Tender offer Bids", *Journal of Financial Economics*, Volume 32, pp. 195-221.
- Chandler, A.D. (1977) *The Visible Hand: The Managerial Revolution in American Business*. Cambridge, MA: Bellknap Press.
- Chhaochharia, V. and Y. Grinstein. (2005) *Corporate Governance and Firm Value - The Impact of The 2002 Governance Rules*. Johnson School Research Paper Series No. 23-06 . Ithaca, NY: Cornell University.
- Cohen, S.S. and G. Boyd. (2000) *Corporate Governance and Globalization*. Aldershot, UK & Northampton, US: Edward Elgar Publishing.
- Cotter, J.F., A. Shivdasani and M. Zenner. (1997) "Do Independent Directors Enhance Target Shareholder Wealth During Tender Offers?", *Journal of Financial Economics*, Volume 43, pp. 195-218.
- Cubbin, J. and D. Leech. (1986) "Growth Versus Profit-Maximization: a Simultaneous Equations Approach to Testing The Marris Model", *Managerial and Decisions Economics*, Volume 7, pp. 123-131.
- Dalton, D.R., C.M. Daily, C.S. Trevis and R. Roengpitya. (2003) "Meta-Analyses of Financial Performance and Equity: Fusion Or Confusion?", *Academy of Management Journal*, Volume 46, pp. 13-26.
- Deci, E. (1972) "The Effects of Contingent and Non-Contingent Rewards and Controls on Intrinsic Motivation", *Organizational Behavior and Human Performance*, Volume 10, pp. 217-229.
- Defusco, R.A., R.R. Johnson and T.S. Zorn. (1990) "The Effect of Executive Stock Option Plans On Stockholders and Bondholders", *The Journal of Finance*, Volume 45, Number 2, pp. 617-627.
- Demsetz, H. (1983) "The Structure of Ownership and The Theory of The Firm", *Journal of Law and Economics*, Volume 26, pp. 375-390.

- Demsetz, H. and K. Lehn. (1985) "The Structure of Corporate Ownership: Causes and Consequences", *Journal of Political Economy*, Volume 93, pp. 1155-1177.
- Donaldson, G. (1984) *Managing Corporate Wealth*. New York: Praeger.
- Donaldson, G. and J.W. Lorsch. (1983) *Decision Making at the Top*. New York: Basic Books.
- Drago, R. and G.T. Garvey. (1998) "Incentives for Helping On The Job: Theory and Evidence", *Journal of Labor Economics*, Volume 16, Number 1, pp. 1-25.
- Easterbrook, F.H. (1984) "Two Agency-Cost Explanations of Dividends", *American Economic Review*, Volume 74, Number 4, pp. 650-659.
- Easterbrook, F.H. (1985) "Insider Trading as An Agency Problem", in John W. Pratt and Richard J. Zeckhauser (Editors), *Principals and Agents: The Structure of Business*. Boston, Mass: Harvard Business School Press, pp. 81-99.
- Eisenhardt, K.M. (1989) "Agency Theory: An Assessment and Review", *The Academy of Management Review*, Volume 14, Number 1, pp. 57-74.
- Fama, E.F. (1980) "Agency Problems and The Theory of The Firm", *Journal of Political Economy*, Volume 88, Number 2, pp. 288-307.
- Fama, E.F. and M.C. Jensen. (1983) "Separation of Ownership and Control", *Journal of Law and Economics*, Volume 26, Number 2, pp. 301-325.
- Frey, B.S. (2003) *Corporate Governance: What Can We Learn From Public Governance?* CLEF, Comparative Law and Economics Forum, University of California at Berkeley, June.
- Giannetti, M. and Y. Koskinen. (2004) *Investor Protection and The Demand for Equity*. Stockholm School of Economics. Working Paper 526.
- Glassman, C.A. and S.A. Rhoades. (1980) "Owner Vs Manager Control Effects On Bank Performance", *Review of Economics and Statistics*, Volume 62, pp. 263-270.
- Gompers, P., J. Ishii and A. Metrick. (2003) "Corporate Governance and Equity Prices", *Quarterly Journal of Economics*, Volume 118, pp. 1-7.
- Grabowski, H.S. and D.S. Mueller. (1972) "Managerial and Stockholder Welfare Models of Firm Expenditures", *Review of Economics and Statistics*, Volume 54, pp. 9-24.
- Grossman, S. and O. Hart. (1980) "Takeover Bids, The Free-Rider Problem, and The Theory of The Corporation", *Bell Journal of Economics*, Volume 11, pp. 42-64.
- Grossman, S. and O. Hart. (1986) "The Costs and Benefits of Ownership: a Theory of Vertical and Lateral Integration", *Journal of Political Economy*, Volume 94, pp. 691-719.
- Hamner, W.C. (1975) "How to Ruin Motivation with Pay", *Compensation & Benefits Review*, Volume 7, Number 3, pp. 17-27.
- Harris, M. and A. Raviv. (1978) "Some Results On Incentive Contracts With Applications to Education and Employment, Health Insurance, and Law Enforcement", *American Economic Review*, Volume 68, Number 1, pp. 20-30.
- Hart, O. (1995) "Corporate Governance: Some Theory and Implications", *Economic Journal*, Volume 105, pp. 678-689.
- Hart, O. and J. Moore. (1995) "Debt and Seniority: An Analysis of The Role of Hard Claims in Constraining Management", *American Economic Review*, Volume 85, pp. 567-585.
- Hartzell, J.C. and L.T. Starks. (2003) "Institutional Investors and Executive Compensation", *Journal of Finance*, Volume 58, pp. 2351-2374.

- Haugen, R.A. and L.W. Senbet. (1981) "Resolving the Agency Problems of External Capital through Options", *Journal of Finance*, Volume 36, pp. 629-648.
- Hanazaki, M., J.W. Shim, T. Souma and Y. Wiwattanakantang. (2004) *Do Large Shareholders Monitor or Collude With Banks in Japan?*, Working Paper. Asian Development Bank Institute. Tokyo: ADBI.
- Hill, C.W.L. and S.A. Snell. (1989) "Effects of Ownership Structure On Corporate Productivity", *Academy of Management Journal*, Volume 32, pp. 25-46.
- Holmstrom, B. (1982) "Moral Hazard in Teams", *The Bell Journal of Economics*, Volume 13, Number 2, pp. 324-340.
- Holmstrom, B. and S. Kaplan. (2003) *The State of US Corporate Governance: What's Right and What's Wrong?* ECGI - Finance Working Paper No. 23/2003. Brussels: European Corporate Governance Institute.
- Jensen, M. and W. Meckling. (1976) "Theory of The Firm: Managerial Behaviour, Agency Costs and Ownership Structure", *Journal of Financial Economics*, Volume 3, Number 4, pp. 305-360.
- Jensen, M.C. (1986) "Agency Costs of Free Cash Flow, Corporate Finance, and Takeovers", *American Economic Review* Volume 76, pp. 323-329.
- Jensen, M.C. (1988) "Takeovers: Their Causes and Consequences", *Journal of Economic Perspectives*, Volume 2, Number 1, pp. 21-48.
- Jensen, M.C. (1989) "The Eclipse of the Public Corporation", *Harvard Business Review*, Volume 67, pp. 61-74.
- Johnson, S., A.P. Boone and E. Friedman. (2000) Corporate Governance in The Asian Financial Crisis, *Journal of Financial Economics*, Volume 58, pp. 141-186.
- Johnson, S., R. La Porta, F. Lopez-De-Silanes and A. Schleifer. (2000) *Tunneling*, NBER Working Paper 7523. Cambridge, Mass: NBER.
- Jones, G.R. and J.E. Butler. (1992) "Managing Internal Corporate Entrepreneurship: An Agency Theory Perspective", *Journal of Management*, Volume 18, Number 4, pp. 733-749.
- ***Kaplan, R.S. (1983) "Comments on Wilson and Jensen", *Accounting Review*, Volume 58, Number 2, pp. 340-346.
- Kock, C.J., J. Santalo and L. Diestre. (2005) *Corporate Governance & The Environment: Bad Discretion, Good Discretion, and Environmental Firm Performance*. IE Working Paper, WPE05-24.
- Kohn, A. (1988) "Incentives Can Be Bad for Business", *INC. Magazine*, pp. 93-94.
- Kornai, J. (1980) *The Economics of Shortage*. Amsterdam, North Holland.
- Kotter, J. P. (1982) *The General Managers*, New York: Free Press.
- La Porta, R., F. Lopez De Silanes, A. Schleifer and R.W. Vishny. (1997) "Legal Determinants of External Finance", *Journal of Finance*. 52(3) 1131-1150.
- La Porta, R., F. Lopez De Silanes and A. Schleifer. (1998) "Law and Finance", *Journal of Political Economy*, Volume 106, pp. 1112-1155.
- La Porta, R., F. Lopez De Silanes and A. Schleifer. (1999) "Corporate Ownership Around the World", *Journal of Finance*. Volume 54, Number 2, pp. 471-517.
- Lawler, E.E. (1971) *Pay and Organizational Effectiveness: A Psychological View*. New York: McGraw-Hill.
- Lazear, E. and S. Rosen. (1981) "Rank-Order Tournaments as Optimum Labor Contracts", *Journal of Political Economy*, Volume 89, pp. 849-864.
- Leland, H. and D. Pyle (1977) "Informational Asymmetries, Financial Structure and

- Financial Intermediation”, *Journal of Finance*, Volume 32, pp. 371-387.
- Lhuillery, S. (2006) *The Impact of Corporate Governance Practices On R&D Intensities On Firms: An Econometric Study On French Largest Companies*. CEMI Working Papers.
- Lincoln, J.R., M.L. Gerlach and P. Takahashi. (1992) “Keiretsu Networks in The Japanese Economy: a Dyad Analysis of Intercompany Ties”, *American Sociological Review*, Volume 57, pp. 561-585.
- Lins, K.V. and F.E. Warnock. (2004) *Corporate Governance and the Shareholder Base*. International Finance Discussion Papers, 816.
- Manne, H.G. (1965) “Mergers and the Market for Corporate Control”, *Journal of Political Economy*, Volume 73, pp. 110-120.
- Marris, R. (1963) “A Model of The “Managerial” Enterprise”, *Quarterly Journal of Economics*, Volume 77, pp. 185-209.
- Masson, R.T. (1971) “Executive Motivations, Earnings, and Consequent Equity Performance”, *Journal of Political Economy*, Volume 79, pp. 1278-1292.
- Mayer, C. (1988) “New Issues in Corporate Finance”, *European Economic Review*, Volume 32, pp. 1167-1183.
- Medoff J.L. and K.G. Abraham. (1980) “Experience, Performance and Earnings”, *Quarterly Journal of Economics*, Volume 95, pp. 703-736.
- Mehran, H. (1995) “Executive Compensation Structure, Ownership and Firm Performance”, *Journal of Financial Economics*, Volume 38, pp. 293-315.
- Mitton, T. (2001) *A Cross-Firm Analysis of The Impact of Corporate Governance on the East Asian Financial Crisis*. New Orleans: AFA.
- Mookherjee, D. (1988) *Competition and Motivation: An Organizational Perspective*. Working Paper, University in Chicago, Business School.
- Mudambi R. and C. Nicosia. (1998) “Ownership Structure and Firm Performance: Evidence from the UK Financial Services Industry”, *Applied Financial Economics*, Volume 8, pp. 175-180.
- Murphy, K.J. (1985) “Corporate Performance and Managerial Remuneration: An Empirical Review”, *Journal of Accounting and Economics*, Volume 7, pp. 11-42.
- Murphy, K.J. (1999) “Executive Compensation”, in Orley C. Ashenfelter and David Card (Editors), *Handbook of Labor Economics*. Amsterdam: Elsevier, pp. 2485-2563.
- Myers, S. and N.S. Majluf. (1984) “Corporate Financing and Investment Decisions When Firms Have Information That Investors Do Not Have”, *Journal of Financial Economics*, Volume 13, pp. 187-221.
- Myers, S.C. (1977) “Determinants of Corporate Borrowing”, *Journal of Financial Economics*, Volume 5, pp. 147-175.
- Nilakant, V. and H. Rao (1994) Agency Theory and Uncertainty in Organizations: An Evaluation, *Organization Studies*, 15(5) 649-672.
- OECD (1998) *Corporate Governance, State-Owned Enterprises and Privatisation*. Paris: OECD.
- OECD. (2001) *Corporate Governance in Asia: A Comparative Perspective*. Paris: OECD.
- Orsburn, J.D., L. Moran., Ed Musselwhite, and Z.H. Zenger. (1990) *Self-Directed Work Teams: The New American Challenge*. Homewood: Irwin.
- Osterloh, M. and B.S. Frey. (2000) “Motivation, Knowledge Transfer, and

- Organizational Form”, *Organization Science*, Volume 11, pp. 538-550.
- Pondy, L. (1969) “Effects of Size, Complexity and Ownership On Administrative Intensity”, *Administrative Science Quarterly*, Volume 14, pp. 46-61.
- Pryke, R. (1981) *The Nationalised Industries: Policies and Performance Since 1968*. Oxford, Martin Robertson.
- Rediker, K.J. and A. Seth. (1995) “Boards of Directors and Substitution Effects of Alternative Governance Mechanisms”, *Strategic Management Journal*, Volume 16, pp. 85-99.
- Ross, S.A. (1979) “Disclosure Regulation in Financial Markets: Implications of Modern Finance Theory and Signalling Theory”, *Issues in Financial Regulation*, Volume 5, pp. 177-202.
- Sanders, W.G. (2001) “Behavioral Responses of CEOs to Stock Ownership and Stock Option Pay”, *Academy of Management Journal*, Volume 44, pp. 477-492.
- Sappington, D.E.M. (1991) “Incentives in Principal-Agent Relationships”, *Journal of Economic Perspectives*, Volume 5, Number 2, pp. 45-66.
- Sharp, D.J. and S.B. Salter (1997) “Project Escalation and Sunk Costs: A Test of the International Generalizability of Agency and Prospect Theories”, *Journal of International Business Studies*, Volume 28, Number 1, pp. 101-121.
- Shavell, S. (1979) “Risk Sharing and Incentives in The Principal and Agent Relationship”, *The Bell Journal of Economics*, Volume 10, Number 1, pp. 55-73.
- Sheard, P. (1996), “Interlocking Shareholding and Corporate Governance”, in M. Aoki and R. Dore, (Editors), *The Japanese Firm: The Sources of Competitive Strength*. Oxford University Press, Oxford, pp.310-49.
- Shleifer, A. and R. Vishny. (1986) “Large Shareholders and Corporate Control”, *Journal of Political Economy*, Volume 94, Number 3, pp. 461-488.
- Shleifer, A. and R. Vishny. (1997) “A Survey of Corporate Governance”, *Journal of Finance*, Volume 52, pp. 737-783.
- Slater, P. (1980) *Wealth Addiction*. New York, Dutton.
- Simon, H.A. (1945) *Administrative Behavior*. New York, Free Press.
- Simon, H.A. (1991) “Organizations and Markets”, *The Journal of Economic Perspectives*, Volume 5, Number 2, pp. 25-44.
- Smith, A. (1776) *The Wealth of Nations*, Cannan Edition. New York, Modern Library, 1937.
- Smith, C.W. and R.L. Watts (1983) *The Structure of Executive Compensation Contracts and The Control of Management*. Unpublished Paper, University of Rochester.
- Stiglitz, J.E. (1987) *Incentives, Cooperation and Risk-Sharing*. New York: Rowan & Littlefield.
- Trice, H.M. and J.M. Beyer. (1984) “Studying Organizational Cultures Through Rites and Ceremonials”, *Academy of Management Review*, Volume 9, pp. 653-659.
- Tsetsekos, G.P. and R.A. Defusco. (1990) “Portfolio Performance, Managerial Ownership and the Size Effect”, *Journal of Portfolio Management*, Volume 16, pp. 33-39.
- Walsh, J.P. and J.K. Seward. (1990) “On the Efficiency of Internal and External Corporate Control Mechanisms”, *Academy of Management Review*, Volume 15, pp. 421-458.
- Williamson, O.E. (1963) “Managerial Discretion and Business Behavior”, *American Economic Review*, Volume 53, pp. 1032-1057.

- Williamson, O.E. (1975) *Markets and Hierarchies: Analysis and Antitrust Implications*. New York, Free Press.
- Williamson, O.E. (1984) "Corporate Governance", *Yale Law Journal*, Volume 93, pp. 1197-1230.
- Williamson, O.E. (1985) *The Economic Institutions of Capitalism*. New York, Free Press.
- Williamson, O.E. (1970) *Corporate Control of Business Behavior*. Englewood Cliffs, NJ: Prentice-Hall.
- World Bank. (1998) *The Business Environment and Corporate Governance: Strengthening Incentives for Private Sector Performance*. Washington D.C.: World Bank.
- Wruck, K.H. (1988) "Equity Ownership Concentration and Firm Value", *Journal of Financial Economics*, Volume 23, pp. 2-28.

Andreas Feidakis
Bank of Greece
Athens, Greece
AFeidakis@bankofgreece.gr

Prison Population: Ethnicity, Class and Gender

Margaret Giles

Introduction

Prison populations are enumerated in terms of both stock and flow as follows. At any one point in time, the number of prisoners (stock) can be recorded. This is generally referred to as the census. Alternatively, the number of persons received into prisons over a period (flow) can be reported. A subset of these persons will be distinct persons received over the same period (flow). For much of the following discussion, the stock figures provide the basis for deriving, *inter alia*, imprisonment and occupancy rates.

In Australia, the adult prisoner population (using prison census figures and excluding juveniles and offenders in psychiatric or police custody) has increased by 43 percent over the decade to 2004. This compares with a 15 percent growth in the Australian adult population over the same period (ABS 2004). Around the world, similar patterns of growth in prisoner populations have been experienced. For example, the number of jail inmates in the US grew by 78 percent between 1990 and 2001 (Levinson 2002: Volume III, p. 1223). In Haiti, the prison population has more than doubled in the three years to 1998 (1,617 in 1995 to 3,766 in 1998) (International Centre for Prison Studies 2005).

Not all prison systems across the globe have shown increased numbers. For example, in Nepal, the prison population has fallen by 5 percent between 1994 (6,200) and 1998 (5,878) (International Centre for Prison Studies 2005). Nor is the pattern of change in prison populations steady in one direction. For example, in Belize, the prison population rose from 617 in 1992 to 1,043 in 1998, then fell to 903 in 2001 (International Centre for

Prison Studies 2005). There are no definitive reasons for these changes; they are the result of a myriad of complementary influences including, *inter alia*, cohort effects, changes to the criminal code and sentencing guidelines, and community tolerance.

The most comprehensive source for correctional facility statistics is the World Prison Brief which is maintained by the International Centre for Prison Studies in King's College, London. Their searchable website is available in English, French, Spanish and Russian and Portuguese. The key statistics that they report are prison populations and rates of imprisonment in total and for juveniles, females and foreigners, percentages of pre-trial detainees and/or remand prisoners, and prison occupancy rates. The sources of the statistics are the agency or agencies with responsibility for correctional services and/or the agency that maintains national statistics.

For example, prison statistics for Australia are taken from an annual publication by the national statistical agency, the Australian Bureau of Statistics. The World Prison Brief advises that, as at June 30 2004 and based on ABS information, Australia had a prison population, including remand prisoners, of 24,171; an imprisonment rate of 120 prisoners per 100,000 of population; and percentages of remand, female and juvenile prisoners of 20.4 percent, 6.9 percent and 0.1 percent, respectively. Other statistics provided for Australia are the percentage of the prison population that is foreign (16.7 percent as at mid-2004); number of prisons (124 in 2004); and an official capacity of 20,503 beds and an occupancy level of 105.9 percent (both as at June 30 2000). Also reported is the trend in imprisonment rates in Australia from 89 in 1992, 96 in 1995, 107 in 1998 to 116 in 2001.

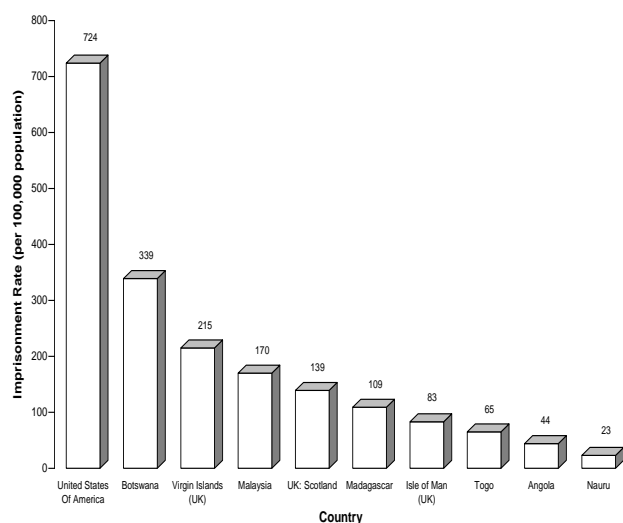
A common way of comparing prison populations for different countries is to look

at imprisonment rates. These rates are calculated from the number of prisoners at a census date (from correctional services/ministry of justice data) per 100,000 head of population at that date (estimated from national data).

The census dates reported in the World Prison Brief for different countries differ. For example, Mozambique is reported with an imprisonment rate of 50 in 1999 and Trinidad and Tobago with an imprisonment rate of 307 in 2003.

Figure 1, below, shows the distribution of imprisonment rates by (approximately) deciles from the highest to the lowest rate. The highest imprisonment rate was in the USA in 2004 with 724 prisoners per 100,000 head of population. The lowest imprisonment rate was 23 prisoners per 100,000 head of population in both Nauru (in 2005) and Burkina Faso (in 2002).

Figure 1. Imprisonment Rates by (Approx) Decile, 10 Nations, 1998 to 2005



Source: Adapted from International Centre for Prison Studies (2005)

Table 1, below, shows high and low imprisonment rates by regional area. Of interest in this table are the high and low imprisonment rates for North America. Both rates are considerably higher than high and low rates in other regions. Interpreting

imprisonment rates is difficult because they reflect a number of jurisdictional peculiarities. These include the specifics of criminal and civil law, the efficacy of policing and judiciary processes in different jurisdictions, and the capacity (both official and applicable) of prison systems. For example, some countries might give custodial sentences for first and minor offences; others are more lenient with sentencing.

Table 1. International Imprisonment Rates^a 2004^b

Region	High	Low
Africa	344 (South Africa)	23 (Burkina Faso)
Caribbean	559 (St Kitts & Nevis)	42 (Haiti)
Central America	470 (Belize)	57 (Guatemala)
Asia	c489 (Turkmenistan)	29 (Nepal)
Europe	569 (Russian Federation)	30 (Faeroe Islands)
Middle East	250 (United Arab Emirates)	61 ^d (Iraq)
North America	724 (USA)	116 (Canada)
Oceania^c	478 (Palau)	23 (Nauru)
South America	437 (Suriname)	74 (Venezuela)

Source: Adapted from International Centre for Prison Studies (2005)

Notes: a. Calculated as number of prisoners per 100,000 head of population. For most countries, this includes pre-trial detainees and remand prisoners; b. Some rates were unavailable for 2004 so figures for the closest year are used; c. The Australian imprisonment rate is 120; d. Includes prisoners in the custody of the Ministry of Justice (the comparative population) as well as prisoners deemed a security threat and held by US occupation authorities.

Another issue with interpreting imprisonment rates is the fact that many prisoners are not serving a prison sentence; they may be awaiting a trial or sentencing (on remand without bail). Table 2, below, highlights proportions of unsentenced prisoners in select countries.

Table 2. Proportions of Unsenteded Prisoners in Prison Populations around the World, 2004^a

Country	Pre-trial detainees and remand prisoners
Andorra	77.0
Australia	20.4
Burkina Faso	58.3
Canada	28.0
France	35.7
Gibraltar	52.6
Maldiv Islands	n.a.
Singapore	8.9
Timor-Leste	70.9
Tuvulu	0.0
UK (England and Wales)	16.9
USA	20.2

Source: Adapted from International Centre for Prison Studies (2005).

Notes: a. Some rates were unavailable for 2004 so figures for the closest year are used.

The growth in prisoner numbers in many countries has not been accompanied by growth in prison facilities. Occupancy rates are often over 100 percent. According to the International Centre for Prison Studies (2005), seventy percent of prison systems were over-utilised in the first years of the new millenium. That is, official prison capacities are exceeded. The highest occupancy rate was recorded in Kenya in 2004 with 3.5 prisoners for every prison bed.

The remaining thirty percent of prison systems have occupancy rates ranging from 99.9 percent for Germany to 27.9 percent for Gibraltar (International Centre for Prison Studies 2005). The low rate for Gibraltar is probably due to economies of scale in prison building resulting in accommodation for 68 prisoners far exceeding prisoner numbers - 19 at the census date in 2004.

In Australia, the occupancy rate is about 106 percent. In Western Australia, the pressure on capacity has seen de-

commissioned accommodation units within some prisons being brought back into service, single cells having one or two extra beds, and new larger prisons replacing smaller prisons. Other sentencing options have also allowed offenders with prison sentences to serve their time under strict conditions in the community.

In the United States, about 2 percent of State prisoners are accommodated in supermaximum security prisons with each providing around 20,000 beds. These supermaxes have been built during a time of hyper-incarceration (see Simon 2000) suggesting an 'economies of scale' mentality by correctional authorities. Leena Kurki and Norval Morris (2001) report that little is known about sentencing requirements for, and conditions in, these prisons. However, they suggest that, contrary to evidence, the "proliferation of supermaxes appears to be premised on a belief that prison disorder is the product primarily of disruptive inmates rather than the characteristics of prison regimes".

Prison populations are not a mirror of wider society. This applies to prisons in first world as well as third world countries. In particular, prisons are populated with proportionately more men, more ethnic minorities and more people from lower socioeconomic backgrounds. The latter also encompasses people with fewer labour market skills and less work experience, and people whose highest levels of educational attainment are well below national averages.

There are a number of reasons why the prison and general population profiles differ. A popular myth is that intelligent criminals don't usually get caught, or, if they do, they don't get caught for a long time or they get off. In that way, more sophisticated (so-called white collar) crimes tend to go unsolved or take decades to solve. Generally, offenders who are caught and charged are a non-

random subset of those who commit offences, and prisoners are a non-random subset of those who are charged. To get into either subset can be determined using the parameters of expected probability theory - the probabilities of being caught, charged and imprisoned together with the likely consequences (one of which is incarceration) of these actions.

Further reasons for prisoners to differ from the general population relate to environment and network effects. Lack of desirable social and professional networks can compound disadvantages for people seeking legitimate activity including work. Not only can lack of desirable networks hinder attempts to locate work but presence of undesirable networks can encourage criminal activity. On the flipside, family and community relationships can be a desirable way of fighting crime (in terms of developing countries, see Carneiro, Loureiro et al 2005).

Race or Ethnicity

The racial mix in prisons can be discussed in terms of natives and foreigners. In Australia, natives are Australian-born indigenous (aborigines and Torres Strait islanders) and non-indigenous people or immigrants who have accepted citizenship and call themselves Australian. Foreigners may be immigrants without citizenship or visitors. All are subject to Australian laws.

The imprisonment of foreigners varies across countries for many reasons. In Australia, drug and people trafficking account for many of the imprisoned foreigners who make up about one sixth of the prison population. Table 3, below, shows that in Andorra, 83.6 percent of the prison population is foreign. This is higher than would be expected with two thirds of the population being immigrants (legal and illegal), mostly from Spain, Portugal, and

France, and attracted to Andorra's zero income tax policy (CIA 2005).

Table 3. Proportions of Foreign Prisoners in Prison Populations around the World, 2004^a

Country	Foreigners
Andorra	83.6
Australia	16.7
Burkina Faso	n.a.
Canada	n.a.
France	21.4
Gibraltar	68.4
Maldiv Islands	n.a.
Singapore	17.6
Timor-Leste	n.a.
Tuvulu	0.0
UK (England and Wales)	12.5
USA	6.5

Source: Adapted from International Centre for Prison Studies (2005).

Notes: a. some rates were unavailable for 2004 so figures for the closest year are used.

Prisoners who fall into the native category can be further classified as part of the population majority group or from a minority group. In Australia, about one quarter of prisoners is indigenous natives. This is considerably higher than their population incidence (about 2.4 percent). In the United States, the imprisonment rates for blacks and Hispanics are higher than their proportions in the population. Numerous hypotheses support these rates, including racial discrimination by majority-controlled justice systems and minority experience of economic and social disadvantage.

Social Class and Education

A number of Australian and overseas studies have examined, inter alia, the prior education, training and work experience of prisoners (Haigler, Harlow et al 1994; Cordella 1995; Germanotta 1995; Thomas 1995; United Nations and UNESCO Institute for Education 1995; Gillis, Motiuk et al. 1998; Greene

1998; Little Hoover Commission 1998; Reuss 1999; ANTA 2000; Hamlyn and Lewis 2000; Petersilia 2000; Worthington, Higgs et al. 2000; ANTA 2001; Lochner and Moretti 2001; Batchelder and Pippert 2002; Department of Justice 2002; Social Exclusion Unit Great Britain 2002; Sutton 2002; WA Department of Justice 2002, 2002; Bearing Point 2003; Burgess 2003; Dormer 2003; NZ Department of Corrections 2003). The main finding from studies between 1992 and 2003 is that a large proportion of the prison population has had minimal schooling and limited prior work experience.

Most of the literature suggests that crime incidence and recidivism are inversely related to the educational attainment and labour market experience of the individual (Kling and Krueger 2001; Batchelder and Pippert 2002; Social Exclusion Unit Great Britain 2002). That is, offenders are more likely to be less educated, have less stable employment histories and have lower incomes than non-offenders. Unemployment and low wages are compounded for repeat offenders due to having a criminal record, possibly including imprisonment. Some of the literature, however, finds that the effects of unemployment and poor levels of education, among other pecuniary and attitudinal variables, on recidivism are either insignificant or sensitive to specification of the variables (Withers (1984) cited in Worthington, Higgs et al. 2000).

Overall, the benefits of reduced recidivism via reduced incarceration costs and lower crime costs are thought to outweigh the costs of the correctional educational programmes. That is, "a significant part of the social return to education comes in the form of externalities from crime reduction" (Lochner and Moretti 2001).

Burgess (2003) has a word of caution for this single focus approach to the provision of education and training in prisons. Specifically

he argues that, in the light of the pursuit of education as a right, targeting recidivism should not be a paramount consideration. That is, there are other social spillovers to the provision of educational services. For example, Sharp et al. (2004) include a better functioning democratic process due to greater voter literacy, more enlightened citizens making society a more pleasant place to live, better government services to the community and more rapid technological process as additional benefits to society.

By encouraging education as a right and not a privilege, correctional authorities are allowing prisoners more choices and giving, particularly young people, the skills to live their lives with structure and purpose (ANTA 2003). This has impacts beyond the prisoners themselves, "reducing the likelihood that their own children will struggle at school" (Social Exclusion Unit Great Britain 2002).

Clearly the focus of education and training in prisons will influence the type of study programmes offered. If the focus is about reduced recidivism and improved labour market outcomes then the education and training programmes are more likely to be vocationally directed. If the focus is on education as a right, then more generalised education programmes could be needed. A broader view would embrace both foci.

The debate about the focus of education and training in prisons, embodied in the above viewpoints, is unlikely to diminish. Indeed it has been fuelled by antagonists of correctional education who claim that justice for victims is not served by taxpayer-funded prisoner education; rather, improving the human capital and therefore the potential lifetime earnings of offenders is an affront to those who lost their lives or were adversely affected—physically, emotionally and financially—by criminal activity (Greene 1998). Thus, the issue is often not about what type of education and training should be

offered in prisons but whether it should be offered at all.

Whilst labour market scholars and the wider community assert the importance of study and work experience, the recognition of the opportunities afforded by such investment in human capital by the prisoners themselves is not clear. For example, there are suggestions that attendance at education courses inside prison “is mainly as a boredom release, not to gain anything specific” (Cook 1990: 97) or “is a means to keeping prisoners occupied” (Batchelder and Pippert 2002; Social Exclusion Unit Great Britain 2002) or “contributes to the ease with which a correctional facility is run” (Batchelder and Pippert 2002:269). However, this finding has been challenged by prison authorities, criminal justice researchers and VET providers (Giles and Le 2004).

Improving opportunities for participation in education, particularly VET, in Australia is seen as a way of addressing inequity in the community. Particular disadvantaged groups have been identified as women, indigenous people, people with disabilities, those with English as a second language, people with inadequate literacy and numeracy skills, and people from rural and remote areas as well as prison inmates (Noonan 2003). The difficulties for achieving equity within prisons relate to the purpose and ramifications of incarceration for the individual, the difficulties of maintaining security, and mental and physical health issues for inmates (including behaviour management needs). Noonan (2003) argues that these difficulties pose challenges for both learners and custodial staff in prisons. They are also a challenge to prison authorities and justice administrators. That is, it is not a simple matter of allowing study and work to be available on site. Recognition needs to be given to other factors in the prisoners' backgrounds, such as their offences, their

substance abuse history and their sense of their future.

Gender

Incarceration for women has generally received different treatment compared to that for men. There are a number of reasons for this, summarised succinctly by Candace Kruttschnitt and Rosemary Gartner (2003). First, the proportion of female inmates in national prison populations is usually low.

Table 4, below, shows the proportions of females in prison populations for select countries. Women make up proportionately less of prison populations than their percentage in the general population. As Pat Carlen and Anne Worrall (2004) summarise “crime tends to be dominated by young men” and “if men behaved like women, the courts would be idle and the prisons empty”.

Table 4. Proportions of Female Prisoners in Prison Populations around the World, 2004^a

Country	Females
Andorra	7.3
Australia	6.9
Burkina Faso	1.0
Canada	5.0
France	3.8
Gibraltar (lowest)	0.0
Maldives Islands (highest)	26.6
Singapore	11.0
Timor-Leste	0.3
Tuvulu	0.0
UK (England and Wales)	6.0
USA	8.7

Source: Adapted from International Centre for Prison Studies (2005). Note: a. some rates were unavailable for 2004 so figures for the closest year are used.

The Maldives Islands have the highest proportion of women in prison populations at 26.6 percent. This may reflect the legal system which is based on Islamic law (tendency for women to be treated more harshly than men for the same offences) or

the rigour of the democracy movement that is pushing reform (women being involved as much as men in actions in support of reform).

Only eleven countries have proportions of females in prison higher than ten percent. The range for the remainder of countries is 9.7 percent for Macau to 0 percent for Gibraltar, Liechtenstein, Nauru, San Marino and Tuvulu.

Another reason for females to receive different treatment inside prison is that their offences tend to be less serious. In their study of adult prisoners in Western Australian metropolitan prisons in 2003, Giles *et al.* (2004) found that females had shorter sentence lengths (70.5 percent had sentences under five years compared with 58.4 percent of males) and lower recidivism rates (about 56 percent were in prison for the first time compared with 45 percent of males).

A final reason for treating women behind bars differently is a belief in their "reformability" (Kruttschnitt and Gartner 2003: 2). This is evidenced in the aforementioned Western Australian study with female prisoners, relative to male prisoners, being more inclined to undertake study in prison and being more likely to expect improved job prospects on their release (Giles *et al.* 2004).

Juveniles

Most justice systems have always treated juveniles (children up to the age of 18 years) differently to adults. This is particularly true of sentencing where imprisonment (juvenile detention) is less likely for juveniles than for adults for the same crime. However, young age in some jurisdictions now no longer provides automatic immunity from the sanctions applied to adults. This has been due to rising rates of youth crime and an inability of current juvenile sanctions to have a deterrent effect.

As a result, there have been two general developments - rising rates of imprisonment and increasing transfers of adolescents from juvenile to criminal courts for adult prosecution. Donna Bishop (2000) suggests that the latter initiative, which reflects a belief in the deterrent effects of criminal sanctions, is ill-placed. She argues that the transfer policy results in sending many minor and non-threatening offenders to the adult system where their special needs may be ignored and their vulnerability may expose them to physical and emotional harm. Bishop also refer to "credible evidence that prosecution and punishment in the adult system increase the likelihood of recidivism, offsetting incapacitative gains" (p. 81). Most jurisdictions are struggling to find alternative sanctions that have the disciplinary force and the necessary deterrent effects to keep young people away from a life crime.

One example of an alternative sentencing option that appears to have merit in practice is used in Western Australia. Here, high rates of apprehension of indigenous juveniles have resulted in non-custodial sentences that involve supervision and mentoring by community elders. Non-custodial penalties tend to be the most frequent sanction for both indigenous and non-indigenous juvenile offenders. Whilst one third of juveniles with burglary/theft offence types and one quarter of juvenile with offences against the person might receive a custodial sentence, between 0.5 and 17.4 percent of juveniles with other offence types are detained (Ferrante *et al.* 2005).

Other alternative options in Western Australia, under the Young Offenders Act 1994, generally require compliance to a general or specific programme and will depend on the nature of the offence and substance-use issues. For serious offences of for juvenile offenders with a long history of being in trouble, sentence to a detention

centre or prison is s worst case scenario. At June 30 2004, 118 juveniles were in detention in Western Australia, representing an imprisonment rate of 51.9 per 100,000 juvenile persons. This is higher than all other States/Territories and twice the national average. The detention rates of indigenous and non-indigenous juveniles were 655 and 13, respectively (Ferrante *et al.* 2005).

Table 5, below, shows proportions of juveniles in prisons in select countries. The countries with the highest and lowest proportions of juvenile prisoners are Tuvalu with 14.3 percent and Andorra and Gibraltar with 0.0 percent. Australia has a low proportion of prisoners who are juveniles, reflecting alternative methods of sentencing such as in home detention and community service.

Table 5. Proportions of Juvenile Prisoners in Prison Populations around the World, 2004^a

Country	Juveniles ^b
Andorra	0.0
Australia	0.1
Burkina Faso	2.4
Canada	10.7
France	1.0
Gibraltar	0.0
Maldives Islands	n.a.
Singapore	7.3
Timor-Leste	2.1
Tuvalu	14.3
UK (England and Wales)	3.1
USA	0.45

Source: Adapted from International Centre for Prison Studies (2005).

Notes: a. some rates were unavailable for 2004 so figures for the closest year are used. b. children aged under 18 years except in Singapore where the age threshold is 21 years.

Issues for the Future

One of the pressing issues for corrections authorities, particularly those whose prison occupancy rates are high, is how to contain the growth in prisoner numbers. Non-

custodial sentencing options are one solution but this may require cooperative community attitudes. Another way to reduce overcrowding in prisons is for case management to better tailor sanctions to offender circumstances. That is, generic sentencing options could be replaced with individualised correctional and rehabilitative programmes.

Another issue for many jurisdictions and their stakeholders is to embrace the rehabilitative model of offender management, including prisons. Recommendations from recent research that highlights the benefits to the community (via reduced recidivism) of in-prison training and work should be adopted and best practice courses and work opportunities made available. In addition, cognitive skills training should be made generally available together with other life skills.

The wider community has a role to play in the future of offender management as well, in particular, in terms of recognising the relative socioeconomic disadvantage that provokes many offenders to crime and antisocial behaviour. Finally, special attention should be given to addressing the needs of ex-prisoners seeking a place in the community (employment, housing, education) after their release.

Further Information

The National Criminal Justice Reference Service (NCJRS), administered by the Office of Justice Programs in the US Department of Justice, maintains a comprehensive list of journals from around the world (www.ncjrs.gov). Topics covered include prison management (such as *The Prison Journal* and *Corrections Management Quarterly*), courts (such as *Justice System Journal*, *Crime to Court* and *Judicature*), sentencing (see *Law and Society Review*), law enforcement (for example, *Policing: An*

International Journal of Police Strategies and Management, *Police Quarterly*, *Canadian Journal of Criminology* and *Journal of Criminal Justice*), crime (see *Trends in Organized Crime*, *Addiction*, *Journal of Family Violence*, *Journal of Financial Crime*, *Journal of Gang Research* and *Journal of Criminal Justice*) and crime prevention (for example, *Crime Prevention and Community Safety: An International Journal*).

Journals for specific crimes (see *Journal of School Violence* and *Journal of Family Violence*) and collectivities (*Journal of Gang Research*, *Juvenile and Family Court Journal* and *Women and Criminal Justice*) are also considered. Justice policy and practice journals include *American Journal of Criminal Justice*, which includes articles on criminal justice processes and the interplay between justice system components and stakeholders, together with policy development, implementation and evaluation issues. The NCJRS list is updated annually and allows title, subject and author searches.

Selected References

- ABS. (2004) *Prisoners in Australia. December*. Cat No. 4517.0. Canberra: ABS.
- Australian National Training Authority. (2000) *National Strategy for VET for Adult Prisoners and Offenders in Australia*. Brisbane: ANTA.
- Australian National Training Authority. (2001) *National Strategy for Vocational Education and Training for Adult Prisoners and Offenders in Australia*. Brisbane: ANTA.
- Australian National Training Authority. (2003) *Shaping Our Future: A Discussion Starter for the Next National Strategy for VET 2004-2010*. Brisbane: ANTA.
- Batchelder, J. S. and J. M. Pippert. (2002) "Hard Time or Idle Time: Factors Affecting Inmate Choices Between Participation in Prison Work and Education Programs", *The Prison Journal*, 82, 2, 269-280.
- Bearing Point. (2003) *Education and Training Provision in Victoria Prisons: The Way Forward*. Melbourne: Office of the Correctional Services Commissioner.
- Burgess, M. (2003) *Post Release and Prison Education Research*. Sydney: A. Webb.
- Carlen, P. and A. Worrall. (2004) *Analysing Women's Imprisonment*. Cullompton: Willan Publishers.
- Carneiro, F.G.; P.R.A. Loureiro and A. Sashsida. (2005) "Crime and Social Interactions: A Developing Country Case Study", *Journal of Socio-Economics*, 34, 3, 311-318.
- Cook, K. (1990) *The Offender's Point of View*. Keeping People out of Prison Conference. Canberra: Australian Institute of Criminology. Papers.
- Cordella, P. (1995) "Prison, Higher Education, and Reintegration: A Communitarian Critique", in H. S. Davidson (Editor), *Schooling in a 'Total Institution': Critical Perspectives on Prison Education*. Westport, Connecticut: Bergin and Garvey, 147-158.
- Criminal Intelligence Agency (CIA) (2005) *The World Fact Book*. Washington DC: CIA.
- Department of Justice (2002) *Profile of Women in Prison*. Western Australia. www.justice.wa.gov.au/content/files/profile_of_women_in_prison.pdf
- Dormer, R. (2003) *Vocational Training in Australian Prisons*. The 9th EPEA International Conference on Prison Education. Norway: EPEA. Papers.
- Ferrante, A.; N. Loh; M. Muller; G. Valuri and J. Fernandez (2005) *Crime and Justice Statistics for Western Australia: 2004*. Perth: Crime Research Centre, University of Western Australia.

www.crc.law.uwa.edu.au/facts_and_figures/statistical_report_2004?f=101638

- Germanotta, D. (1995) "Prison Education: A Contextual Analysis" in H.S. Davidson (Editor), *Schooling in a 'Total Institution': Critical Perspectives on Prison Education*. Westport, Connecticut: Bergin and Garvey, 103-121.
- Giles, M. and A.T. Le. (2004) *Labour Market Activities of Prisoners*. Paper presented to the Australian Labour Market Research Conference, University of Western Australia, December 6-7.
- Giles, M.; A.T. Le; M. Allan; C. Lees; A.-C. Larsen and L. Bennett. (2004) *To Train or Not to Train: The Role of Education and Training in Prison to Work Transitions*. Perth: Centre for Labour Market Research. www.ncver.edu.au/research/proj/nr3022b.pdf
- Gillis, C.A.; L.L. Motiuk and R. Belcourt. (1998) *Prison Work Program (CORCAN) Participation: Post-Release Employment and Recidivism*. Ottawa: Correctional Service Canada.
- Greene, M.K. (1998) "'Show Me the Money': Should Taxpayer Funds be Used to Educate Prisoners Under the Guise of Reducing Recidivism", *New England Journal on Criminal and Civil Confinement*, 24, Winter.
- Haigler, K.O.; C. Harlow; P. O'Connor and A. Campbell. (1994) *Literacy Behind Prison Walls*. US National Centre for Educational Statistics. Washington DC: USGPO.
- Hamlyn, B. and D. Lewis. (2000) *Women Prisoners: A Survey of Their Work and Training Experiences in Custody and on Release*. London: Home Office. www.homeoffice.gov.uk/rds/pdfs/hors208.pdf
- International Centre for Prison Studies. (2008) *World Prison Brief Online*. King's College. www.kcl.ac.uk/depsta/law/research/icps/worldbrief
- Kling, J.R. and A.B. Krueger (2001) "Costs, Benefits and Distributional Consequences of Inmate Labor", Princeton University Industrial Relations Section Working Paper No. 449.
- Kruttschnitt, C. and R. Gartner. (2003) "Women's Imprisonment", in M. Tonry (Editor), *Crime and Justice*. Chicago: University of Chicago Press, 30: 1-81.
- Kurki, L. and N. Morris. (2001) "The Purpose, Practices, and Problems of Supermax Prisons" in M. Tonry (Editor), *Crime and Justice*. Chicago: University of Chicago Press, Vol. 28.
- Levinson, D. (2002) (Editor) *Encyclopaedia of Crime and Punishment*. Thousand Oaks, CA: Sage Publications.
- Little Hoover Commission (1998) *Beyond Bars: Correctional Reforms to Lower Prison Costs and Reduce Crimes*. Report No. 144. Sacramento, California, LHC. www.lhc.ca.gov/lhcdir/144/TC144.html
- Lochner, L. and E. Moretti (2001) *The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-reports*. NBER Working Paper 8605. New York: NBER.
- Noonan, P. (2003) *Equity in Education and Training in Correctional Services Institutions*. Adelaide: National Centre for Vocational Research.
- New Zealand Department of Corrections (2003) *Census of Prison Inmates and Home Detainees 2001*. Auckland: Department of Corrections.
- Petersilia, J. (2000) "When Prisoners Return to the Community: Political, Economic, and Social Consequences", *Sentencing and Corrections: Issues for the 21st Century*, 9, November, 1-7.
- Reuss, A. (1999) "Prison(er) Education." *The Howard Journal*, 38, 2, 113-127.
- Sharp, A.M.; C.A. Register and Paul Grimes. (2004) *Economics of Social Issues*.

Sixteenth Edition. Homewood, Illinois: US, Irwin.

Simon, J. (2000) The 'Society of Captives' in the Era of Hyper-Incarceration. *Theoretical Criminology* 4: 285-308.

Social Exclusion Unit Great Britain. (2002) *Reducing Re-offending by Ex-prisoners*. London: UK Government.

Sutton, J.R. (2002) *Imprisonment and Labor Market Outcomes: Evidence from 15 Affluent Western Democracies*. Department of Sociology, UCSB. www.soc.ucsb.edu/faculty/sutton/Design/Assets/Compris4.pdf

Thomas, J. (1995) "The Ironies of Prison Education" in H. S. Davidson (Editor), *Schooling in a "Total Institution": Critical Perspectives on Prison Education*. Westport, Connecticut: Bergin and Garvey, 25-41.

Tonry, M. (2001, 2003) (Editor) *Crime and Justice: A Review of Research*. Chicago, University of Chicago Press, Vols. 28, 30.

United Nations and UNESCO Institute for Education (1995) *Basic Education in Prisons*. Vienna and Hamburg: IOE.

WA Department of Justice (2002) *Adult Education in WA Prisons: an Overview of the Past 100 Years*. Perth: WA Department of Justice.

WA Department of Justice (2002) *Profile of Women in Prison*. Perth: WA Department of Justice.

Worthington, A., H. Higgs, *et al.* (2000) "Determinants of Recidivism in Paroled Queensland Prisoners: A Comparative Analysis of Custodial and Socioeconomic Characteristics", *Australian Economic Papers*, September.

Websites

International Centre for Prison Studies. www.kcl.ac.uk/schools/law/research/icps

Prisoners in Australia, 1996-2006. www.aic.gov.au/publications/cfi/cfi147.html

US Criminal Justice Reference Service
www.ncjrs.gov/whatsncjrs.html
World Prison Population List.
www.homeoffice.gov.uk/rds/pdfs2/r188.pdf

Margaret Giles
School of Business
Edith Cowan University
Perth, Australia
m.giles@ecu.edu.au

Prostitution

Johanna Kantola

Introduction

In everyday language, prostitution refers to the exchange of sex or sexual services for money or other material benefits. In academic terms, it can be defined as a social institution that allows certain powers of command over one person's body to be exercised by another (O'Connell Davison 1998:9). Most prostitution involves heterosexual sexual exchanges, with men buying the sexual services from women usually in the context of unequal power relations between the sexes (Outshoorn 2003).

Prostitution is a hotly-debated issue. It is a concern for a wide range of actors, from sex workers and feminists, to politicians, different state bodies, international organisations and neighbourhood associations. Controversy and arguments abound on all aspects of prostitution. What kind of problem is prostitution? How should prostitution be dealt with? Sex workers, in contrast, want to highlight prostitutes' problems rather than prostitution as a problem. States and international organisations are increasingly confronted with a need to deal with prostitution and their solutions depend on the ways in which they frame prostitution. Currently, countries are taking diverging directions in their prostitution policies. While The Netherlands and Germany have legalised prostitution, Sweden has criminalised buying sex. Many debates on prostitution are now being dominated by concerns for trafficking in women and children.

The Problem of Prostitution

Notably often prostitution is defined in gender-neutral ways where prostitution is framed as a law and order problem, public nuisance, a moral problem, and a public

sexual health problem (Phoenix 1999, Outshoorn 2003). As a *law-and-order problem* prostitution is thought to give rise to rowdiness and drunken behaviour. Prostitution is seen to relate to other criminal activities, such as drugs, money laundering and violence. In the *public nuisance* frame, prostitution is thought to create an environment where property values and business in red light districts decline through the 'bad reputation' of that area, where 'innocent and decent men' are accosted, where road traffic is disrupted by kerb crawlers, where fear of molestation is increased by the presence of men on the streets late at night, and where young children are exposed to the manifestations of prostitutions (Phoenix 1999:22).

In the *moral problem* frame, prostitution is seen as a sin or a vice and it is a manifestation of deviant sexuality. The prostitutes – 'fallen women' – should be saved and redeemed. Also state regulation is targeted because the state is seen as a 'pimp' and facilitating prostitution (Outshoorn 2003). In the 19th Century, the Contagious Diseases Acts legally formalised the construction of prostitution as a *public sexual health problem*. The Acts enabled police and state interventions into the lives of prostitutes in the name of medical inspection (Walkowitz 1982). When prostitution is framed as a public sexual health problem, prostitutes are constructed as sexually unclean and diseased others (Phoenix 1999:29). The conceptualisation was consolidated in the 1980s and 1990s with the arrival of the AIDS epidemic.

For feminists, prostitution has been a difficult and divisive issue. Nevertheless, feminists have shared a need to insert new gender meanings of women, men and sexual practices into the debates (Outshoorn 2003). Feminist analyses on prostitution linger between two positions. On the one hand,

feminists such as Kathleen Barry (1995) and Sheila Jeffreys (1997) define prostitution as *sexual domination* and as the essence of women's oppression. On the other hand, a number of feminist authors now maintain that prostitution is work that women can opt for (Pheterson 1989, Chapkis 1997, Sullivan 1997). In the *sex work perspective*, prostitution is seen as an income-generating activity or form of labour for women and men, and prostitute women's agency is emphasised (Kempadoo 1998:9).

Prostitutes' Problems

The feminist sex work perspective has been able to theorise prostitutes' problems as opposed to solely seeing prostitution as a problem. Prostitutes started to organise in the 1970s to articulate their views. Early examples included Call Off Your Old Tired Ethics (COYOTE) in San Francisco (1973), French Collective of Prostitutes (1975), English Collective of Prostitutes (1975), New York Prostitutes Collective (1979) that later became USPROS, Australian Prostitutes Collective (1981), now known as the Prostitutes Collective of Victoria (PCV), The Dutch Red Thread and HYDRA in Germany. The International Committee for Prostitutes' Rights (ICPR) was established in 1985 and two World Whores Congresses were held, one in Amsterdam in 1985 and one in Belgium in 1986.

Key struggles for sex workers and feminist theorists supporting them involve recognition of women's work, basic human rights and decent working conditions. Specific concerns include sexual harassment, family and personal leave, pay, job evaluation, breaks, and dismissals (English Collective of Prostitutes 1997).

Both world congresses were dominated by Western agendas despite the fact that prostitute organising did exist in the so-called Third World countries (Kempadoo 1998).

Examples include the Ecuadorian Association of Autonomous Women Workers (1982) and Uruguayan, Association of Public Prostitutes (AMEPU) (1985). Neither was an integral part of the 'international' movement. In the beginning of the 1990s, the Network of Sex Work Projects (NSWP) began to address this problem. The Network draws upon the sex work discourse and has now brought together 40 different groups from all parts of the world.

The Network for Sex Work Projects stresses the problems of racism and postcolonialism in the sex work industry. It is not simply grinding poverty that underpins a woman's involvement in prostitution. Race and ethnicity are equally important factors for any understanding of contemporary sex industries (Kempadoo 1998:10). Even with the heightened exoticisation of the sexuality of Third World women and men, they are positioned within the global sex industry second to white women and work in poorest and most dangerous sectors of the trade, particularly street work.

Prostitution Policies

Prostitution policies are usually categorised in terms of abolition, prohibition or regulation, or decriminalisation and legalisation. *Abolitionism* refers to the position that argues that prostitution should be banned and third parties criminalised, with the prostitute herself not liable to state penalties. Those drawing upon the moral problem frame and the feminist sexual domination frame promote this policy. *Prohibitionism*, makes all prostitution illegal and all parties liable to penalties, including the prostitute. *Regulation* is an overall term denoting state intervention in the running of prostitution. Policies range from allowing brothels or red light districts to different degrees of control over prostitutes (Outshoorn 2003; West 2000). Law-and-order, public nuisance and public sexual

health problem frames advocate different forms of regulation. *Decriminalisation* aims to normalise prostitution, removing the social exclusion which makes prostitutes vulnerable to exploitation, and sex work then becomes subject to regulation by civil employment law. *Legalisation* implies state regulation through licensing or registration and compulsory health checks, with outlets or workers not granted permits still subject to criminal penalties (West 2000). Sex work perspectives argue for decriminalisation or legalisation. Notably, one nation state's measures are not internally homogenous and may include elements of more than one regulatory framework.

Australia and The Netherlands have opted for legalisation of prostitution. In Australia, street walking or public soliciting for the purposes of prostitution is legal in New South Wales. Brothels and/or escort agencies may operate openly in Victoria, New South Wales, Queensland, the Northern Territory and the Australian Capital Territory. In most cases, brothel premises and the owners and operators of brothels (not workers) are subject to licensing and sex workers employed in legal prostitution businesses have many of the same rights as other Australian workers (Sullivan 1997).

National legislation regarding prostitution in The Netherlands used to be abolitionist and anyone involved in the organisation of, or living of, prostitution was criminalised. However, since this legislation was introduced in 1911, there has been a gradual relaxation in implementation which has meant that prostitution has, in reality, long been accepted as a way of life (Kilvington et al 2001:81). A new law came into effect in 2000 and the Dutch Penal Code no longer treats organising the prostitution of an adult female or male person as a crime, provided this is done with the consent of the prostitute. If a woman regards prostitution the best way

to earn a living, she has the same rights as any other worker. Any form of forced prostitution, pimping or trafficking, remains in the Penal Code, with the maximum of six years of imprisonment. Prostitution is managed within a system of licences, where several municipal services are responsible for checking the conditions under which prostitution is allowed to operate. Conditions include city planning (no brothels are permitted near schools or churches), state of the buildings (sanitary and safety issues), and finally, management (no forced drinking, no unsafe sex, no minors, no illegal workers, the owner must not have a criminal record) (van Doorninck 2002).

Arguments for legalisation are manifold. Prostitution will become more open, easier to control for inspectors and more professional. People with criminal records seeking to make quick profits will not be allowed to become a part of the legal sex industry. Insurance companies and banks can no longer discriminate against sex workers which will have an empowering effect on the women who are used to having no rights at all (van Doorninck 2002; Kilvington et al 2001). Arguments against legalisation stress the position of migrant workers who fall outside of the new legislation. For example, in The Netherlands, the position of migrant women from outside the EU worsened, and illegal sex workers moved across the border to Belgium and Luxembourg (van Doorninck 2002; Kilvington et al 2001:86). Feminists, who are for abolition of prostitution, argue that in Australia legalisation has led to an increase in crime and women have become more vulnerable (Sullivan & Jeffreys 2002).

Britain and Sweden represent different approaches in comparison to Australia or The Netherlands. In Britain, the public nuisance frame has dominated prostitution debates. The British approach could be characterised as one of negative regulation (Phoenix 1999).

The law has not sought to abolish or legally repress prostitution by criminalising the sale of sexual services, but neither has the law been used to regulate prostitution by legalising it. Campaigns against kerb crawling have been the most visible manifestation of the British prostitution policies and the law has been changed to first introduce the offence of 'persistent kerb crawling' (Sexual Offences Act 1985) and then kerb crawling was made an arrestable offence (Criminal Justice and Police Act 2001). The police aims to reduce or end street prostitution but tolerate the growth of less visible forms of prostitution such as sauna work, brothel work and home work, which are seen as causing less public nuisance.

Sweden, in turn, has criminalised the purchase of prostitutes' services in the sex-buying law in 1998. The law is underpinned by the idea that all prostitution is a social problem and women are seen as victims rather than as criminals or sex workers. The aim of the law was to reduce the number of women in prostitution. However, this effect has been debatable and arguably the law has only led to a decrease in the visibility of prostitution. Prostitutes have been pushed underground, which has exposed them to violent acts and to working for less money. It has also become increasingly difficult for project workers in Sweden to contact sex workers for support, advice, information and counselling (Kilvington et al 2001:85). Feminists in Sweden argue, however, that the law has send an important signal to the society:degrading women by buying sex is not acceptable behaviour in Sweden (Gould 2001).

Trafficking in Women or Transnational Prostitution Migration?

Concerns for trafficking in women and children now dominate debates on prostitution in many countries. However, sex

work across national boundaries is not a new phenomenon. It emerged as a political issue in the 1880s, when it was called 'white slavery' (Walkowitz 1982). It is also hard to evaluate whether the activity has intensified because it is difficult, if not impossible, to obtain reliable statistics on the phenomenon. Nevertheless, a number of trends is exacerbating the position of women worldwide:*feminization of poverty*, gendered effects of structural adjustment programmes (cuts in government social expenditure and rise in neo-liberalism), women as docile labour force (seasonal and flexible employment), feminisation of international labour migration, increase in sex industry and in sex tourism, and state support to commodification of women.

Trafficking and migration trends flow from Eastern Europe and West Africa to Western Europe, Thai women work in Germany and in Canada, Asian sex workers in Australia. Women are trafficked or they migrate from Nepal to India, and Bangladesh to Pakistan. Many of these women are aware that they will be working in the sex industry and decide to go in the belief that they will be able to make substantial amounts of money.

International law has had only limited tools to deal with the illegal and coerced aspects of trafficking in women. Attempts include the 1979 Convention on the Elimination of All Forms of Discrimination Against Women (CEDAW), the Vienna Declaration on the elimination of violence against women (1993) and the UN Beijing Conference's *Platform for Action* (1995). The Declaration condemned forced prostitution and trafficking, but not prostitution *per se*; the Platform for Action called for fighting forced prostitution. In 2000, the UN Protocol on Trafficking was finally agreed on, and it formed a comprehensive international attempt to stop trafficking. It defined trafficking as the recruitment and transfer of persons by

means of the threat or use of force or coercion, fraud, deception or abuse of power for the purpose of exploitation. Exploitation includes the prostitution of others, sexual exploitation, forced labour or services, slavery, servitude or the removal of organs. The Protocol required state parties to penalise trafficking and to protect victims of trafficking and grant them temporary or permanent residence in the countries of destination (Sullivan 2003).

Intensive debating took place around the Protocol, and two positions can be discerned. The Coalition Against Trafficking in Women (CATW) founded by Kathleen Barry in the US in 1991 represents the first perspective. CATW defines prostitution as a form of sexual exploitation just like rape, genital mutilation, incest and domestic violence. The second perspective, represented by the Global Alliance Against Trafficking in Women (GAATW) based in Thailand, objects to forced prostitution but wants to separate it from voluntary prostitution. The prostitutes' rights movement developed the distinction between voluntary and forced prostitution in response to feminists and others who saw all prostitution as abusive (Doezema 1998:37). The Global Alliance has been moving towards a position where trafficked women are treated as migrant labourers and they are protected by international labour legislation. Feminists drawing upon this perspective argue that the trafficking protocol is unable to protect the rights of the voluntary undocumented migrant women in the sex trade. Abuse of human rights is more likely to occur as a result of the work being illegal and informal than through deceptive coercion and lack of consent (Thorbek & Pattanaik 2002).

Barry's work and CATW's portrayals of victims of trafficking have captured the imagination of policy makers on international and national levels, parts of the media and the public. This has masked the underlying

colonial and postcolonial tendencies that underpin the analyses and racist ideas about 'developing countries' and 'Third World women'. Barry argues that 'trafficking' could involve any woman in the world, and that any woman could become a sex slave. However, a closer reading reveals that she constructs a hierarchy of stages of patriarchy and development, situating trafficking in women in the 'pre-industrial and feudal societies' where women are exclusively the property of men. At the other end of the spectrum are the developed countries where prostitution is normalised and where women are independent and emancipated (Kempadoo 1999:11). Her analysis reproduces Western ideologies and moralities regarding sexual relations. Third World women are seen to be trapped in underdeveloped states and Third World prostitutes continue to be positioned in this discourse as incapable of making decisions about their own lives, forced by overwhelming external powers completely beyond their control into submission and slavery. Western women's experience is made synonymous with assumptions about the inherent superiority of industrialised capitalist development and Third World women placed in categories of pre-technological 'backwardness', inferiority, dependency and ignorance (Kempadoo 1999:12).

Assessment of Prostitution

The manner in which trafficking in women has come to dominate debates on prostitution in many countries is dangerous. It is narrowing all debates on prostitution to coercion and forced prostitution and is thus silencing sex workers' voices. It also results in disregard of other important aspects of prostitution. Important areas, where more research is needed, include male prostitution, child prostitution and research on customers. Also homosexual and transgender sex work

are understudied areas in the literature (see however Weitzer 2000).

The narrowing of all prostitution debates into trafficking has resulted in some activists and academics arguing for separating the two issues. In such a framework, we can perhaps distinguish two types of strategies to combat specifically trafficking in women. On the one hand, repressive strategies argue for more restrictive immigration policies, penalisation and stronger and more effective prosecution. These have a strong tendency to end up working against women rather than in their favour and yet, they are the most attractive to governments as they fit to various 'tough on crime' agendas. On the other hand, there are strategies that aim to strengthen the rights of the women involved, as women, as female migrants, as female migrant workers and as female migrant sex workers. These support strategies aim to empower women, in contrast to stigmatising or victimising them (Wijers 1998:77-78). As such, the strategies are not separate from the more general picture of prostitution where sex workers would benefit from a change in public attitudes towards prostitute women, from an emphasis on their rights to decide over their own destinies and sexualities. In other words, strategies that support women in sex work without stigmatising and victimising them would be effective in making the position of prostitutes, migrant sex workers and trafficked women better.

Selected References

- Barry, Kathleen. (1995) *The Prostitution of Sexuality: The Global Exploitation of Women*. New York: New York University Press.
- Chapkis, Wendy. (1997) *Live Sex Acts: Women Performing Erotic Labor*. New York: Routledge.
- Davis, Nannette J. (1993) (Editor) *Prostitution: An International Handbook on Trends, Problems and Policies*. Westport, CT: Greenwood Press.
- Doezema, Jo. (1998) "Forced to Choose: Beyond the Voluntary vs. Forced Prostitution Dichotomy", in Kamala Kempadoo and Jo Doezema (Editor), *Global Sex Workers: Rights, Resistance and Redefinition*. New York & London: Routledge, pp 34-50.
- van Doorninck, Marieke. (2002) "A Business Like Any Other? Managing the Sex Industry in the Netherlands", in Susanne Thorbek and Bandana Pattanaik (Editors), *Transnational Prostitution: Changing Global Patterns*. London and New York, Zed Books, pp 193-200.
- English Collective of Prostitutes. (1997) "Campaigning for Legal Change" in G. Scambler and A. Scambler (Editors), *Rethinking Prostitution Now*, London, Routledge, pp 83-103.
- Gould, Arthur. (2001) "The Criminalisation of Buying Sex: the Politics of Prostitution in Sweden", *Journal of Social Policy* Volume 30, Number 3, pp 437-456.
- Jeffreys, Sheila. (1997) *The Idea of Prostitution*. North Melbourne: Spinifex.
- Kempadoo, Kamala. (1998) "Introduction: Globalizing Sex Workers' Rights", in Kamala Kempadoo and Jo Doezema (Editors), *Global Sex Workers: Rights, Resistance and Redefinition*. New York & London, Routledge, pp 1-28.
- Kilvington, Judith, Sophie Day, and Helen Ward. (2001) "Prostitution Policy in Europe: A Time of Change?", *Feminist Review*, Volume 67, pp 78-93.
- O'Connell Davidson, Julia. (1998) *Prostitution, Power and Freedom*. Cambridge: Polity Press.
- Outshoorn, Joyce. (2003) "Introduction: Prostitution, Women's Movements and Democratic Politics", in Joyce Outshoorn (Editor), *The Politics of Prostitution: Women's Movements, Democratic States*

- and the Globalisation of Sex Commerce*. Cambridge, Cambridge University Press.
- Pheterson, Gail. (1989) *A Vindication of the Rights of the Whores*. Seattle, WA: The Seal Press.
- Phoenix, Joanna. (1999) *Making Sense of Prostitution*. London, Macmillan.
- Sullivan, Barbara. (1997) *The Politics of Sex. Prostitution and Pornography in Australia since 1945*. Sydney, Allen and Unwin.
- Sullivan, Barbara. (2003) "Trafficking in Women: Feminism and New International Law", *International Feminist Journal of Politics*, 5, 1, 67-91.
- Sullivan, Mary Lucille and Sheila Jeffreys. (2002) "Legalization: The Australian Experience", *Violence Against Women*, 8, 9, 1140-1148.
- Thorbek, Susanne and Bandana Pattanaik. (2002) (Editors), *Transnational Prostitution: Changing Global Patterns*. London and New York: Zed Books.
- Walkowitz, Judith. (1982) *Prostitution and Victorian Society: Women, Class and the State*. Cambridge: Cambridge University Press.
- Weitzer, Ronald. (2000) (Editor) *Sex for Sale: Prostitution, Pornography and the Sex Industry*. New York and London: Routledge.
- West, Jackie. (2000) "Prostitution: Collectives and the Politics of Regulation", *Gender, Work and Organization*, 7, 2, 106-118.
- Wijers, Marjan. (1998) "Women, Labour and Migration: The Position of Trafficked Women and Strategies for Support", in Kamala Kempadoo and Jo Doezema (Editors), *Global Sex Workers: Rights, Resistance and Redefinition*. New York & London: Routledge, pp 1-28.
- Wijers, Marjan. (2000) "European Union Policies on Trafficking in Women" in M. Rossilli (Editor), *Gender Policies in the*

European Union. Oxford: Lang, pp 193-208.

Websites

- Coalition Against Trafficking in Women. (CATW) www.catwinternational.org
- Network of Sex Work Projects (NSWP) www.nswp.org/home.html
- Global Prostitution Resources. arapaho.nsuok.edu/~dreveskr/prolinks.html
- [l-ssi](#)

Johanna Kantola
 Department of Political Science
 University of Helsinki
 Helsinki, Finland
johanna.kantola@helsinki.fi

Research and Development

Jerry Courvisanos

Introduction

Research and development (R&D) is the most organised form of innovative activity. It sets out to create, apply and diffuse new knowledge in a structured process: It embodies a process “whereby new and improved products, processes, materials, and services are developed and transferred to a plant and/or market. Typically, this process is represented in the firm by a number of formally organized laboratories, departments, groups, teams and functions...involv[ing] scientists and engineers” (Burgelman *et al* 1996:2) Rosenberg (1982:120) sees this as “a learning process” in the generation of new technical knowledge.

Early R&D involved corporate in-house learning, which all major corporations set up after the Second World War, whether in the form of the ubiquitous laboratory for manufacturing or more diversely as ‘new product development’ within the marketing department. In the services sector, the locus of learning activities occurs often in groups called ‘business development’ or ‘technology’. Smaller firms also have R&D activities appearing under the titles of ‘design’ or ‘technical support’, but rarely a specific plant facility or business unit. All the above require the exchange of information across organisational boundaries within the firm in a ‘closed innovation’ system that carries with it a large proportion of firm-specific culture and knowledge.

Since the early 1990s, R&D has increasingly involved an ‘open innovation’ system through a distributed innovation process that leverages knowledge from a broad variety of sources outside the firm itself, including university research, contracting research from ‘centres of

excellence’, joint venture consortiums, acquired entrepreneurial firms and licensing of innovations. Thus, boundaries for firms conducting R&D have broadened widely under cost pressures and the evolution of the internet with its supporting web-enabling technologies. “Increasingly firms are acknowledging that it is difficult for them to create and exploit technological innovations on their own.” (Bowonder *et al* 2005:51) Large firms in high-tech sectors where there is rapid technological change occurring, like in semiconductors, show greater propensity for embracing the open innovation system and relying on young small creative R&D firms to keep abreast with the frontiers of knowledge (Miotti & Sachwald 2003).

Research is scientific or technological investigation that has the potential to lead to an idea or concept for innovation. This research is conducted usually by specific experts in two different stages. The basic research stage is exploratory with no preconceived outcome or direction, and no clearly identified practical applications, but needs to have present or potential interest to the organisation conducting the investigation. This research is associated with scientific discovery or, more generally, knowledge-building. The applied research stage has preconceived goals based on business imperatives related to specific products, processes or service delivery. This research is problem-solving and needs to take basic research into practical applications that have indefinable private and/or social returns that relate to the strategic positioning of the organisation.

Development explores the specific potential of a product, process or service within an experimental testing environment. This work needs to be conducted between the technical experts, logistical production managers and marketing departments. Two stages can be identified. First is the blueprints

stage, where a set of designs for specific outcomes are developed from theoretical research. This is followed by the prototype stage, which creates test models for technical feasibility. In all stages of ‘development’ there is need for continual feedback to R in order to improve the theory. This is an iterative process with many failures and dead ends along the way, but is essentially a linear model of R&D innovation.

The first major researcher on innovation, Joseph Schumpeter, developed two frameworks on innovation that can be used to appreciate R&D. In Schumpeter’s first analysis in 1911, he identified the entrepreneurial process in terms of the small capitalist who drives new ideas into the market place while destroying old products and processes (“creative destruction”, Schumpeter, 1934), and this seemed to be consistent with the form of capitalism observed by economists through the 19th Century. The innovative activity is seen to be exogenous to the firm (especially the characteristics of the entrepreneur), in what has been referred to as Schumpeter Mark I. There is no official R&D undertaken under Mark I, as small entrepreneurs develop, test and market their ideas in the process of producing and selling the idea. At the time, this analysis ignored the nascent rise of the R&D process within corporations instituted in USA by Thomas Edison through his R&D laboratory and factory set up in Menlo Park, New Jersey in 1876, and around the same time in Germany (Freeman & Soete 1997).

By the early 1940s, Schumpeter recognised the institutionalisation of R&D in sustaining the monopoly power of large corporations to the point that he was concerned that this process would see the end of the entrepreneur, as R&D becomes a purely bureaucratic activity (Schumpeter, 1942). This raised the spectre of Schumpeter Mark II with “creative accumulation” from

minor incremental innovative activity that is endogenous to the large corporation. Courvisanos (2005a) explains the open system interaction of the Mark I and II processes as advanced capitalism moves into the 21st Century, where the small innovative firm complements the R&D process within large firms. Many of the new ideas refined in the R&D process are identified and initially developed by small firms who are closer to the customer and the market place. The large firms set up their R&D through collaborations with, or acquisitions of, small firms.

Investment

R&D expenditure is often referred to as spending, yet conceptually these funds are knowledge-based investment, analogous to capital investment. Even failed R&D projects contribute to the corpus of knowledge by identifying what does not work and creates further problems to be solved with more investigations. In this way, R&D spending is a significant part of what economists now call ‘intangible investment’ because it is an investment into future production, but the knowledge-base that such investment creates is not tangible and obvious as plant and equipment (capital) investment (Webster 1999). Since January 1 2005 company balance sheets in Europe can include all such intangibles as assets (in forms such as patents, trademarks, and even firm-specific knowledge) at ‘fair market’ prices.

Given the broad non-specific nature of basic research, it seems surprising that so many firms invest in it, since the published and publicly available research outcomes can be appropriated by anyone. Rosenberg (1990) dismisses this “information view” because it undervalues the embodied knowledge required to benefit from basic research outcomes. Firms invest in basic research in order to effectively understand, evaluate, monitor and network into the broad scientific

community and its knowledge base. Cohen and Levinthal (1989) develop a model of how firm investment in R&D “creates the capacity to assimilate and exploit new knowledge” (p.593). Further, Rosenberg (1990) argues that the first-mover advantages gained from such an approach far outweighs the costs of undertaking basic research. Basic research information remains tacit until there is an investment in embodying it in firm-specific knowledge (Salter & Martin 2001:512).

The various stages of R&D identified above all have different outputs, each one is difficult to evaluate and has diverse possible outcomes, with greatest uncertainty at the early stages of the R&D process. This renders R&D for radical innovation highly problematic (Hall 2002:36-7). Thus, evaluation depends on the judgement of experts at the various R&D stages, both inside and outside the firm. However, Mansfield *et al.* (1972) in a classic study identifies that these experts tend (when they are planning) to greatly underestimate development costs, while they greatly overestimate the time taken to produce results. Then, Tidd *et al.* (2005:218) note that scientists and engineers in basic and applied ‘R’ are often deliberately overoptimistic in their estimates in order to give the illusion of a high rate of return to conservative accountants and managers. As a result, R&D management requires a much more effective communication between R&D staff and the persons responsible for allocation of financial resources, as well as seeking outside advice on the management of the R&D portfolio (Ettlie 2000:149).

The Mansfield study of project selection in large US firms, comparing forecasts to outcomes found that the probability of picking winners by these firms was only 16 per cent. Jolly (1997), 25 years later, confirms that the Mansfield results still stand even with the advances in modern computer technology.

Thus, attempts to manage market and technical uncertainty generally fail. This leads large firms to invest profits in R&D to protect successful innovations in order to maximise returns over as long a period as possible. Protection can be legal (e.g. secrecy and patents) or illegal (e.g. cartel arrangements), creating monopoly power for that period of protection.

Measurement

R&D is measured in three ways; input, output and capacity. From the process context there is a set of input measures of R&D, notably: R&D expenditure in absolute monetary figures, R&D intensity (ratio of R&D expenditure to sales), number of R&D employees, and R&D stock. R&D expenditure is a common *absolute* measure of financial commitment, while R&D intensity takes this spending figure and provides a *relative* measure in relation to sales. Employee numbers gives some notion of human resource capability, while technological capability is better measured by the R&D stock built up in the organisation. Large manufacturing sectors like automobiles and electronics are strongly represented in the absolute input measure, while rising industries like biotechnology are represented strongly in the relative input measure.

For the output of creative and intellectual effort, there is a different set of R&D measures. Most common in marketing terms is the number of sales of new (up to 5 years after launch) products relative to total annual sales. R&D productivity measured by income from new products relative to R&D expenditure (lagged 3-5 years) is a favourite of the accountants who allocate R&D funds. From a more technical perspective there is R&D output intensity measured by the number of patent applications to real (deflated) R&D investment. At a purely scientific level, there is the number of

scientific articles published in quality science journals and the extent that these articles are cited over the following few years. Finally, technology licences issued to other firms is an indication of the extent of diffusion of the R&D innovation, but this is generally viable as a measure only where the technology can be easily ‘unbundled’ and adopted by other firms with different institutions and culture.

Building technological capacity for effective R&D can also be measured, but by more qualitative indicators from inside and outside the firm. Internal indicators identify the extent of technical expertise, focus on end-user needs, cross-functional and fluid research teams, strategic focus and formal development processes (Menke 1997). External indicators identify specific capacity-building R&D structures; especially lead user groups (von Hippel 2005), innovation networks, research consortia, strategic alliances and joint ventures related to market developments (Tidd *et al.* 2005:296-315). These R&D activities can be strongly identified with growth industries like pharmaceuticals and telecommunications arising chiefly out of the expansion of external sources of R&D under the distributed innovation process (Bowonder 2005:51-2).

The most difficult to measure is R&D done *not* under the banner of ‘R&D’. Particularly, this is the case in two areas. One is ‘informal R&D’ undertaken in less organised and *ad hoc* basis (e.g. trouble shooting on the production line). In many countries, especially large ones, informal R&D is regarded as too difficult to survey (Pavitt 1994), whereas some small countries like Australia survey all size firms for all forms of R&D (Bryant 1998:59, fn.4). The other is service-based and some product-based efforts in innovation which use electronic information technology for investigation and testing. Rosenberg (1982:191) identified this problem a long

time ago when he said: “Software development shares many of the problems of any R&D activity.” Freeman (1994) noted the rise of information and communications technology (ICT) as an area of innovation itself did not come out of identifiable R&D activities. Bowonder *et al.* (2002) identify the emergence of e-engineering and e-design for innovation as central to R&D but not measured in current R&D metrics.

Innovation and Firm Size

There is a major debate over the issue of R&D as the source of significant technological innovation and its location within the size of firms. Source of creativity by R&D is a two-edged sword. The technology imperative demands two conflicting actions. Firms need to maximise gains from any successful innovation developed in-house, while focusing on radical innovations for a distinctive competitive edge. Tension exists between the two scenarios for R&D, incremental or radical innovation: Incremental represents minor improvements on existing products or processes that require little organisational change, while radical innovation represents revolutionary departure from current operations with significantly different skills and capabilities.

Abernathy and Utterback (1985) provide a technology life-cycle model that can explain the dual role of R&D. Product innovation creates new goods and services. If these creations are significantly different from the products that replaced them, then they are ‘radical’ and will elicit, from within the firm, significant R&D-induced process innovations so to reduce production costs. Eventually as product matures, the stimulus for process innovation fades as well, and only incremental innovations emerge at a decreasing rate. Incremental innovation provides ‘extra’ profits from successfully

application of radical innovations with only marginal R&D input. Financial managers encourage further incremental innovation, since it can be calibrated easier with simple 'rules-of-thumb' for allocating resources, establishing sunset criteria for projects and using sensitivity analysis based on a known range of assumptions while reducing key uncertainties before commitment (Tidd *et al.* 2005:218-20). This limits the innovative edge as financial controllers seek short-term gains.

'Incrementalism' is further entrenched by marketing efforts and monopoly power. Professional R&D executives recognise the role of marketing in its interaction with lead users of the products in setting R&D agendas. This is done not only by standard marketing 'research' surveys, but increasingly more prevalent has become collaboration with lead users on finding what such users 'need' to improve use of their products, e.g. mountain bikes and computer software (von Hippel 2005). Monopoly power of secrecy and property rights aim to stifle radical product innovation being conducted by smaller entrepreneurial firms or even in-house radical ideas which threaten the current strong market position of the dominant firm(s). Lessig (2004:3-7) provides evidence of this from the media industry. For example, the RCA company squashed all attempts by their R&D engineer, Edwin Armstrong, to introduce the higher quality FM radio band; all RCA wanted was to protect their radio monopoly by reducing the static noise on the AM band. Further, incrementalism by defensive publishing to protect patents has become a significant alternative strategy to R&D investment (Bar 2006).

Radical (or major) innovation has its limitations at both ends of the corporate world. Very small firms are not able to invest in R&D, thus there is a minimum size threshold before firms can be considered R&D-based. The inherent creative power

within small innovative R&D firms is undermined by the well documented failings of venture capital markets to support highly uncertain radical innovation. The lack of realised profits by these young start-up R&D-based firms means that they depend greatly on external sources of funding, especially from venture capitalists (Tassey 1997:190). Although Freeman and Soete (1997) acknowledge the evidence that smaller firms appear disproportionately among inventors and patentees, their R&D efficiency is severely constrained by finances and limited by other resources like skilled labour.

At the other end are the large firms that have the advantage of many projects and functions, allowing complementarities and cost spreading to provide higher returns on R&D investment (Cohen and Klepper 1996). This allows larger firms to dominate major process innovation in a drive for productivity increases, while acquiring many small firms' major product innovations to complete the commercialising process (Legge 2000). However, three factors associated with large firms, being bureaucracy, uncertainty and monopoly power, undermine radical outcomes. It is for this reason that Scherer (1980) identifies increases in R&D investment with reduced number of radical innovations.

Ettlie and Rubenstein (1987) examination of 348 US manufacturing firms identifies smaller firms (up to 1,000 employees) as introducing radical and incremental new products at nearly the same rate, then as firms increase their size up to 11,000 employees their greater size tends to promote more radicalness. When firms become very large (greater than 11,000 employees), there is a clear trend to 'incrementalism' and a lack of radical product innovation despite often having very large R&D units. (See also supporting evidence for Canada in Baldwin 1997).

The simple conclusion on firm size and innovation suggested by Pavitt *et al.* (1987) and empirical supported by Tsai and Wang (2005), is that there is a 'U-type' relationship between R&D productivity and firm size, with medium-sized firms having the disadvantages of both small and large firms and without the advantages of either. This supports the Mark I and II approach, but with interaction between the two being crucial (Baldwin 1997). R&D effort is enhanced if the interaction is supportive (as small firms' product innovations are developed further by large firms and added to with process innovation), or is inferior if the interaction is obstructive (as small firms' patents are bought by large firms and not developed).

Radical innovations are significantly more likely to be commercially successful (Ettlie and Rubenstein 1987). This is because the accumulated firm-specific intangible knowledge for future opportunities (first mover advantage) tends to be greater the more radical the innovation. The difficulty is assessing which ideas will eventually succeed and having to pursue many on the expectations that one will succeed. There is fundamental uncertainty with no probability distribution, and thus no calculable risk assessment that can be made for successful radical innovations. Ettlie (2000:40) estimates that only 6-10 per cent of all new successful products are radical, while successful radical processes are even scarcer. Focusing on radical innovations will not only require a considerable shift in skill capability and organisational structure, but it also introduces the threat of new entrants (some very large with 'deep pockets') into the industry who are prepared to diffuse the innovation. Concentration on radical R&D requires brave foresight on the part of established business.

Up to the early 1960s R&D was funded directly from central corporate sources. Since then there has been a growing movement to

fund from contracts between the R&D division and other internal and external business 'groups', e.g. Philips (with 5 labs around the world) began in 1990 to have its funding from head office reduced to one-third, with the remainder coming from contracts from business groups (Jolly 1997:346). This trend threatens creative R&D in radical innovations and tends to support incremental innovation driven by contract-based strategic marketing needs. Philips, realising this, modified their funding structure in 1994, requiring roughly half of the two-thirds controlled by business groups to be "devoted to immediate product development; the remaining half has to be for longer-term capability development in certain technology clusters, such as signal processing for TVs. Typically, this part is funded by more than one business group as well." (Jolly 1997:349)

The dilemma is determining how R&D strategies address both short-term market-based needs and long-term knowledge-accumulation needs. This all depends on the valuation of strategic intent by the firm undertaking R&D. The technology life-cycle model assists in appreciating the nature of this strategic intent. During the growth stage of a successful innovation, incremental changes out of R&D result in substantial gains for the firm and in terms of social benefit as the innovation is adapted and diffused. Then as the innovation matures, R&D tends to suffer diminishing returns in terms of new knowledge and new applications. At this mature stage defensive R&D efforts aim to maintain market position (Bar 2006) and create significant barriers to new radical product innovations coming out of the R&D efforts of competing small and large firms. This also creates major barriers in the R&D divisions of defensive-oriented firms if they want to shift R&D operations to becoming offensive. (See Freeman & Soete 1997; the IBM example when it set up a

complete newly staffed R&D division in order to become offensive again.)

Business Evidence

Empirical evidence on the role of business R&D in innovation is multi-faceted and inconsistent. The business sector performs 68 per cent of total R&D in the OECD countries and this sector is the major source of financing domestic R&D, accounting for almost 62 per cent of OECD funding in 2003. Business R&D funding varies sharply across the OECD, with 75 per cent of R&D in Japan being funded by the business sector in 2003, 63 per cent in the USA and down to 55 per cent in the 15 member EU where publicly-funded R&D is much stronger. Business R&D in real terms has steadily increased since 1980, with the USA increasing by 3.2 per cent p.a. between 1995 and 2003, EU15 by 3.7 and Japan by 3.5 (OECD 2005).

Despite the value of investing in basic research identified above, Whiteley (1994) finds that only four per cent of the 1992 allocation of R&D funds by members of the US Industrial Research Institute was spent on basic research (down from six per cent in 1988). Compare this to 41 per cent spent on new product development and one can clearly see the priorities of large research intensive firms back in 1992 and this broad picture remains a constant. Small and medium sized enterprises (SMEs) with fewer than 250 employees play an important role in entrepreneurial innovation and competitive pressure for large firms, but only account for around 30 per cent of total R&D expenditure. In the USA there is much 'small business' rhetoric yet SMEs account for less than 15 per cent of business R&D while in Japan it is nine. New Zealand (72), Norway (70), Ireland and Greece (49) and the Slovak Republic (46) are the only countries to have significant business R&D accounted for by SMEs (OECD 2005).

Bowonder *et al.* (2005) report on global firms shows the highest number of patents per million dollars of R&D spending in their respective industry are large and established, not the market sales leaders but "close on their heels". Firms with the highest R&D intensity in their respective industry generally also have a significantly smaller sales base, indicating that they are "up-and-comers" whose R&D efforts have not (yet) produced many patents, with higher sales turnover emerging much later on in their technology life-cycle. Exceptions to these patterns are very few. Only Alcatel in telecommunications is that industry's biggest R&D spender as well as having the largest number of patents per million dollars R&D spending (0.173) and the highest R&D intensity (12.42). Only Ajinomoto (in food processing) and Nortel Networks are both the patent industry leaders and have the highest industry R&D intensity. Microsoft is a good example of a market leader whose R&D intensity is 21.12 (higher than the industry average, 13.12), but much lower than BMC Software with an R&D intensity of 39.00 but sales of only \$US1.5 m. (cf. Microsoft on \$US36.8m.). The patent numbers for Microsoft is very low (0.088) compared to their R&D spending (\$US7.779m. in 2004 and forecast around \$US10,000m. in 2006 of the total industry R&D of \$US26.5m.). Agilent Technologies has the software patent leadership at 0.706.

The extraordinary strong R&D by the pharmaceutical industry is led by European-based Sanofi-Aventis (SA) which recorded a 96.7 per cent R&D spending rise between 2003 and 2004. SA continues to be the fastest growing R&D spender, rising to overall lead in R&D spending with a relatively strong R&D intensity of around 16 per cent, and forecasted to spend \$US15,400m. in 2006. This strong R&D growth is a spur to new discoveries and attempts to protect expiring patents (Schonfeld Associates 2005).

Bowonder *et al.* (2005) also examine the acquisition of knowledge outside of internal R&D sources and found this was an increasing trend in all industries except software. The number and share of patents assigned to more than one firm is growing consistently through alliances. Intense competition induces firms to acquire knowledge and technologies from a variety of sources, leading to increasing use of the distributed innovation processes where knowledge creation and the leveraging from outside sources is carefully balanced. For a good example of this, see the SEMATECH case study (Ettlie 2000:164-6).

Public Policies

Central to all nations' *industrial policy* is the approach governments adopt to funding and supporting R&D. Gerschenkron (1962) associates this with the 'late' industrial development stage of the global economy. Economics literature has identified four rationales for such emphasis on supporting what is essentially a private sector activity. First is the neoclassical supply-oriented concern arising from 'market failure', based on inadequate return for the private sector in R&D due to few and uncertain pay-offs from basic research (Arrow, 1962). As noted, large firm R&D tends to support incrementalism. There is limited market-based encouragement for more uncertain radical innovations arising from less powerful industries and firms where the scale of R&D is too low to generate the critical mass of new knowledge. Also, duplication by competitors tends to quickly undermine any competitive edge established by the initiator (free rider issue). From the nation's standpoint, these problems of market failure lead to underinvestment in non-incremental R&D.

The second rationale centres on national security issues developed by Gansler (1980). Ability to be self-sufficient in circumstances

of secrecy on defence (and space program) strategies drives this concern (offensive). It is bolstered by concerns of being cut-off or refusal to trade during military conflicts (defensive). R&D spending, due to secrecy and lack of direct civilian applicability, can not be supported in private markets. This R&D is financed by the public sector, but developed in the private sector, with the use of procurements to drive down R&D costs. In the long-run the knowledge gained provides a platform for new civilian capabilities far into the future (e.g. computers, GPS, commercial space travel). This has been the case throughout history, but clearly at different rates of civilian uptake (White 2005).

The third rationale is based on evolutionary economics, centred on systems approach that rejects the linear model of R&D innovation. The national innovation system is a set of institutions whose complex interaction via clusters, collaborations and networks across the public-private sector space determine the extent of innovative performance (Nelson 1992). In this system, R&D forms the foundation of knowledge and its applicability for innovation. However, systemic failures exist in private sector R&D due to lock-in, transitional problems, poor knowledge-based *infrastructure*, and inappropriate conventions and institutions (Smith 1998). For example, private sector R&D support for small firms with single innovation ideas are hard to justify on financial grounds because the chances of success before any patents expire are very low (Legge & Hindle 2004:337). Such systemic failures justify the need for national governments to intervene in a strategic way to complement business R&D. David *et al.* (2000:527) survey the evidence on public R&D and conclude that "[c]omplementarity appears more prevalent, and substitution effects all but vanish among the subgroup of

studies that have investigated this relationship at the industry and national economy level.”

The final rationale is based on environmental concerns. Exhaustion of non-renewable resources and pollution threaten the environment’s ecosystem viability, while markets do not reflect the ecological value of sustainability of human and other life on this planet. Thus, there is a need for public finance and support of R&D on decentralised alternative new energy sources and reducing pollution (McDaniel 2002:85). Neoclassical and evolutionary economists could claim this argument for their respective market or systemic failure arguments, however ecological economists see the ecosystem overriding both such approaches. A market failure approach can merely encourage the public support of R&D into costly and unsustainable ‘end-of-pipe’ technological solutions. A systemic failure approach to work from this environmental perspective needs R&D that has clear ecological directions and rules that allow for adaptation and incremental change towards a decentralised sustainable ecosystem (rather than support, for example, of massive centralised nuclear power and corporate genetic engineering, see Skea 1994).

Two types of R&D public policies are possible, passive and active (Legge and Hindle 2004:237-50). Passive policies respect *laissez-faire* market solutions by attempting to override market failures, giving markets a better chance to work effectively. This would involve *intellectual property rights* (IPRs) protection of R&D innovation to overcome the free rider issue, and providing broad R&D rebates, subsidies and incentives in order to reduce uncertainty and support scale economies. This is the neoclassical approach to R&D public policies. Cannon (2005) explains that the USA, as the leader in R&D, has strong preference for passive R&D policies and notes the four successful R&D

instruments are (in order of importance): tax relief, defence support, patent protection, and college education. The paradox of passivity by not picking winners and yet supporting massive defence R&D does not seem to be apparent in Cannon’s analysis, but this is to be expected from the perspective of the dominant approach that inhabits R&D in the USA. A more recent variation of this neoclassical “passive” approach has been policies to shift R&D support from large corporations towards small business (through programs for technology start-up companies like pre-seed funding and incubators). Though the conservatives could suggest this change is due to market failure as large corporations override the market, such a *post hoc* rationale undermines the whole passive approach and leads directly to active policies.

Active policies aim to directly intervene in order to influence the direction and extent of R&D innovation. Sectoral R&D assistance to specific industries aims to address concerns of the lack of innovation in this area (e.g. CSIRO as the Australian public research body in support essentially of the agricultural sector). Selective public investment in research infrastructure (e.g., synchrotrons, technology parks, cooperative university-business research centres), subsidies in specific areas of concern (environment, social groups, non-urban regions) and public sector procurement of R&D (as in defence industry) all provide direction as part of public policy support. All rationales bar neoclassical tend to support such active policies, with the particular direction of R&D support for political debate (centralised authority or democratic grassroots). The proponents of such active policies argue on the basis that these are emerging areas of economic activity that need support to overcome systemic failures described above.

In reality, R&D public policies end up being a mix of both passive and active,

depending on the political trajectory that a nation has traversed over the last fifty years. The trend of R&D policies will reflect the rationale which is being championed by the political powers at the time. There are, however, some theoretical limitations to R&D support by the state. In relation to subsidy/incentive-type support, successful R&D innovations end up benefiting the private sector firm involved twice, once from financial support and second from profits of the innovation often with state-endorsed monopoly control through IPRs. Concern also exists as producers of R&D get exclusive benefits of the IPRs, when often it is users who generate the innovative ideas but all benefits go to producer who also gets patent (and other) IPR protection (von Hippel 2005). Dolfsma (2006:339) identifies the public policy concern for IPRs “being hijacked by larger firms, particularly for strategic purposes”.

Questions are also raised about governments’ attempts to ‘boost’ R&D when it is used merely as a marketing tool for incremental innovation (How many blades can you place on razor shaver?). This is supported by evidence that incumbent enterprises, with minor innovative activity, benefit most from such R&D public support during long economic expansions; whereas new firm start-ups are triggered by economic contraction and unemployment supported by university research in particular (Audretsch and Acs 1994). At the other extreme with radical innovation, there is the growing neo-liberal influence in many western economies to encourage support for small-based entrepreneurial start-ups based on some “exaggerated claims of their role in innovation” (Legge & Hindle 2004:247), when in fact the vast majority of entrepreneurial start-ups are extensions of work conducted prior to start-up (Bhidé 2000). Further, Åsterbo (2003) shows

evidence of unrealistic optimism in a sample of 1,091 independent inventions, with only between 7-9 per cent reaching the market and 60 per cent of them obtaining negative returns.

Empirical evidence on R&D support is mixed. Bloom *et al.* (2002) draws the conclusion from a nine major OECD-country study that generally R&D tax credits have had a significant effect. However, other studies have found several problems with this form of passive incentive: criteria are stringent, applies only to new R&D, no distinction between R&D spending and success rates, productivity effects are varied, and ignoring increasingly important role of collaborations (Ettlie 2000:298-300).

In terms of active public R&D policies, three nations—Ireland, New Zealand, Australia—have more than half of public financed R&D actively directed to firms with fewer than 50 employees, rare indeed in other countries. The USA, UK, France (as well as some smaller countries like Turkey) have most public-financed business R&D directed to large firms (OECD 2005). Active policies like selective investment (e.g. energy), incubators and technology parks have had varying success, depending on how well targeted the policy is, how well it is administered and monitored; then there is the level of synergy of companies involved with similar and complementary endowments. Finally there is the motivation of the participants themselves in these research infrastructures. Australia has been notable for selective investment in two major successful innovation-based research infrastructures: CSIRO and AIS (Australian Institute of Sport). Both have been models that have been studied and copied around the world, however, Australia’s natural and cultural endowment in agriculture and sport have much to do with motivation and this success (Fox 2001).

Major national issues provide impetus for R&D public policies and their success. Major security concerns ensure that the military rationale will inevitably be successful (even if the military campaign fails). The Netherlands has also been able to develop successful innovation policies for sustainable development on the back of major environmental crises that the populace as a whole recognise and accept (Courvisanos 2005b).

A third of OECD countries (all small economies) have public as their major source of R&D funding, also all less developed economies depend on government for R&D. From this it can be noted that higher education and government sectors perform almost 30 per cent of all R&D (OECD 2005).

Globalisation

R&D spending is gradually globalising in concert with general business globalisation, but not in a straight direct transfer of operations. The evidence, based on country of origin in front page patent citing and company R&D spending data, is quite clear that the majority of R&D by global corporations occurs in their home of origin. Only 12 per cent of the world's large firms conduct R&D outside their home country, compared to around 25 per cent equivalent share of production. Notably, on average foreign-based production is less innovation-intensive than home production, with firms from smaller countries generally having higher shares of foreign innovative activities. Most R&D performed outside home sites occurs in USA and Germany, with a growing trend in biotechnology and ICT for European firms to conduct R&D in USA so as to access local skills and knowledge (Tidd *et al* 2005:211-13). Meyer-Krahmer and Reger (1999) characterise R&D as being in the dominant "Triadization" structure, involving companies from the USA, the EU and Japan.

OECD (2005) figures indicate at the broader level of 30 leading economies (excluding China and India), that well over 16 per cent of total R&D expenditure is performed abroad by foreign affiliates. The picture that emerges is a complex mosaic of rising internationalisation of R&D but with limited "techno-globalism".

Domestic country of origin R&D still matters to large corporations. Tidd *et al.* (2005:213-16) have identified reasons for this: When launching substantially complex new products and processes, there are major efficiency gains from close proximity to R&D developments for knowledge integration and dealing with unforeseen problems. Despite IT linkages, tacit knowledge through close personal contact matters. There are very high fixed costs in setting up such R&D infrastructure outside of the domestic country or region where R&D originally developed and thus created a strong culture and technological trajectory. Spread of R&D depends on the ability of industries to overcome these inherent domestic advantages. Compare an industry's need for R&D to be close to markets and the specific production for these markets (e.g. global firms like Ford produce cars for Australian conditions which need to be different to Europe and USA) with another industry which needs to be close to basic research knowledge from particular knowledge-intensive centres (e.g. global pharmaceutical firms located close to leading universities with pharmaceutical research expertise).

Matching foreign localisation with highly specialised R&D personnel can be difficult and requires significant ability in mobilising such staff. Thus, R&D works best at foreign centres with more established products and services that have moved considerably down the life-cycle for market-seeking ('capability exploiting', CBE) motives, while embryonic ones stay at home. China's strong economic

growth at the beginning of the 21st Century, has encouraged foreign firms to relocate R&D facilities essentially at this CBE level (Dahlman and Aubert 2001:121-38). China is the third most attractive location for foreign R&D, India the sixth. Both offer around one-eighth the R&D costs of OECD nations, but concerns remain that CBE facilities will not lead to significant build up of technological capabilities—as occurred with the Taiwanese experience prior to such facilities being shifted to mainland China (Altenburg *et al.* 2006). This concern is supported by evidence that talented émigrés are difficult to lure back home (Cervantes & Guellec 2002).

In attempts to overcome the limitations of R&D-direct firm globalisation, outsourcing and collaborative R&D on a more global basis have become a strong trend. The agency for this move has been the development of global knowledge networks across the private-public sector space for resource-seeking motives. Scientists and engineers were the first to develop electronic-based global knowledge networks in military, space and then university research, all public sector funded. More recently, business firms have found it very useful to tap into these existing knowledge networks and to extend them further in what Kuemmerle (1997) calls “capability augmenting” (CBA) R&D facilities close to public centres of research excellence. European firms tend to follow this CBA approach strongly, with Ambos (2005) providing evidence to support this with respect to 134 R&D laboratories of top German global firms. These networks link researchers in advanced economies, leaving the rest of the world outside these networks.

Outsourcing of non-core R&D activities in incremental innovation has been the major form of internationalisation with the links being vertical to suppliers and customers (especially lead users) throughout their extended global value chain. Examples of

such outsourcing partners who can reduce transaction costs are systems integrators, technology consultants as well as more traditional suppliers (inputs) and customers (marketing). Increased global sourcing and marketing has allowed for more extensive and flexible outsourcing arrangements, with loose coupling of multi-technology products allowing for uneven rates of advance to be accommodated up and down the value chain (Brusconi *et al.* 2002).

An extensive literature in collaborative R&D identifies a variety of forms, from simple joint firm cooperation, to competitors R&D consortia, to virtual (electronic) collaborations (see Ettlie 2000:159-69). They tend to be associated by way of horizontal links with competitors and work better when stimulated to join for reasons such as start-up phase, threat of new entries, or concern that maturity has set in. Almost all of the 80 per cent growth in technology collaborations since the mid-1980s is accounted for by the high technology areas of pharmaceuticals, biotechnology and ICT who are particularly keen to establish flexible collaborations to allow for technology switching. Aerospace and defence actually have declined their collaborative R&D efforts over the same period (Tidd *et al.* 2005:318-9). Unfortunately, the skill base and political regimes in Africa are not conducive to becoming part of any such potential collaboration. Latin America has some potential; however in Brazil the government is supporting the development of a strong generic pharmaceutical industry to the frustration of the major global pharmaceutical firms (Cohen 2000).

Comparative Economies Evidence

OECD (2005) provides a useful overview, as at 2003, of how successful different countries are in the ‘R&D stakes’. R&D intensity (as % of GDP) has Israel with the highest at 4.9,

then Sweden at 4 – both countries having a strong defence R&D program, Finland, Japan and Iceland are all at 3, with the OECD average at 2.2. China is the third largest R&D spender (10 per cent of global R&D) due to the rapid growth in researcher's salaries, behind USA (35) and Japan (14). US companies have contributed by expanding their R&D spending in China from \$US7 million in 1994 to \$US506 million in 2000. At the purely scientific level, the highest relative intensity of articles per population comes from Sweden, Switzerland and Finland. Since 1995, nations with the fastest growing R&D expenditure in real terms have been Iceland, Turkey and Portugal (all above ten per cent).

As reported in OECD (2005), USA dominated R&D expenditure with a massive 42.1 per cent of total OECD, followed by EU (15 members) at 29.6 per cent (with Germany 8.3, France 5.5, UK 4.7, Italy 2.7, Spain 1.5, The Netherlands 1.3). Outside this group, the other notable economies are South Korea with 3.6, Canada 2.8, Sweden 1.6 and Australia 1.4. The rest of OECD economies have one per cent or less. Outside the OECD only two nations matter, China and India; China quintupling its budget over ten years since the mid-1990s to \$US 84.6 billion and India at \$US20.7 billion is greater than Canada (Altenburg *et al.* 2006: 4). Despite this growth, China's R&D is seven times less than the OECD average as a percentage of value added (Gilboy 2004).

OCED (2005) also reports that Government R&D budgets have increased annually by an average of 3.5 per cent in real terms since 2000 for the whole of the OECD. Three-quarters of the growth in public R&D in the USA between 2001 and 2005 is attributable to defence R&D. A third of OECD countries (all small economies) have public as their major source of R&D funding, also all less developed economies depend on

government for R&D. From this it can be noted that higher education and government sectors perform almost 30 per cent of all R&D. In terms of R&D support, 50 per cent more OECD countries since 1996 provide R&D concessions to firms, rising to 18 economies in 2005. Canada, the Netherlands and Italy focus on small firms, while other economies do not distinguish by size. High-tech manufacturing accounted for more than 54 per cent of all OECD R&D, with 60 per cent of it in USA alone.

Services are poorly measured. One notable figure is the unique difference in R&D growth rates between the services and manufacturing sectors in Ireland. For the period 1993-2001, Irish R&D increased by 27 per cent in services (mainly in computers), but only seven per cent in manufacturing (OECD 2005).

Ethical Issues

Ethics is an important issue in R&D that has been ignored in the past by the scientific and business R&D communities in the technology-push mentality that all research and new knowledge is inherently 'good'. Philosophers and ethists like Peter Singer who question this inherent goodness have been seen as 'cranks', allowed to simply voice their minority view in a democracy. In the 21st Century, with R&D biotechnology enabling the alteration of human *genetics*, the ethical issues of R&D have been rammed into the front of the R&D community. There are now thousands of patents worldwide (except in France) on the human genome, with a global alliance human genome project aimed to increase this number dramatically. The ethics and economics of such R&D are in conflict, and the low standards in modern business ethics indicate that corporate decision-making will not produce the philosophically desirable outcome. Merely survival of businesses and research staff may

not be the best approach for ensuring considered long-term implications of altering the world's gene pool and the ecosystem (Flowers 1998).

Conclusion

Two dominant elements in R&D are national security concerns and incrementalism. National security concerns shape innovation systems and particularly the R&D component. With national security concerns raised even higher after 9/11, there is even greater publicly funded support for defence/security based R&D across all economies, but dominated by the USA and UK as they lead the anti-terrorism strategy. As well, the monopoly power of large corporations to dominate R&D in marketing-based incremental innovation remains.

Two aspects of the R&D system that remain, but of less importance, are in-house R&D and capability exploiting foreign-based R&D activities in cheaper developing economies where there are some strong R&D-based skills. India, for example, has a strong medical training tradition which is being used by the major pharmaceutical companies to assist in undertaking R&D drug trials, but with much reduced ethical standards that have enraged civil rights defenders.

A new development in R&D systems is related to services-based areas of knowledge; much of it not conducted in official R&D centres. Firms in the new economies of India and China are emerging out of technology transfer into their own capability augmenting R&D often in concert with collaborative firm partners from established R&D strong economies. This provides only a very limited globalisation trend in respect to R&D. Lead users and other democratising elements of R&D activities are becoming more established in western economies, providing opportunities for a broader constituency in

developing innovation and challenging the dominant incrementalism. This requires more public policy input into the creativity and commercialisation aspects of R&D as well as greater public debate on the ethics and ecological impacts of R&D. From this there needs to emerge a broader collaborative R&D innovation process that includes alliances and network sharing across a large and diverse range of communities interested and affected by R&D; from conservation groups and trade union bodies, to scientists, corporate leaders and entrepreneurs. However, these emerging features will find it difficult to survive and grow in the face of the dominant security and monopoly power elements of 'vertical' based specialist silos of knowledge.

Selected References

- Abernathy, W.J. and Utterback, J.M. (1978) "Patterns of Technology Innovation", *Technology Review*, 80, 7, 40-7.
- Altenburg, T.; H. Schmitz and A. Stamm. (2006) *Building Knowledge-based Competitive Advantages in China and India: Lessons and Consequences for Other Developing Countries*. paper presented at Global Development Network Annual Conference. St. Petersburg: 18-19 January.
- Ambos, B. (2005) "Foreign Direct Investment in Industrial Research and Development: A Study of German MNCs", *Research Policy*, 34, 4, 395-410.
- Arrow, K. (1962) "Economic Welfare and the Allocation of Resources for Invention", in Nelson, R. (ed.) *The Rate and Direction of Inventive Activity: Economic and Social Factors*, Princeton, NJ: Princeton University Press, 602-25.
- Åsterbo, T. (2003) "The Return to Independent Invention: Evidence of Unrealistic Optimism, Risk Seeking or Skewness Loving?", *The Economic Journal*, 113, 226-39.

- Audretsch, D. and Z. Acs. (1994) "Entrepreneurial Activity, Innovation and Macroeconomic Fluctuations", in Shionaya, Y. and Perlman, M. (Editors) *Innovation in Technology, Industries and Institutions: Studies in Schumpeterian Perspectives*, Ann Arbor: The University of Michigan Press, 173-83.
- Baldwin, J.R. (1997) *The Importance of Research and Development for Innovation in Small and Large Canadian Manufacturing Firms*. Statistics Canada, Working Paper No. 107, Ottawa, 24 September.
- Bar, T. (2006) "Defensive Publications in an R&D Race", *Journal of Economics & Management Strategy*, 15, 1, 229-54.
- Bhidé, A. V. (2000) *The Origin and Evolution of New Businesses*. Oxford: Oxford University Press.
- Bloom, N.; R. Griffith and J. Van Reenen. (2002) "Do R&D Tax Credits Work? Evidence from a Panel of Countries 1979-97", *Journal of Public Economics*, 85, 1, 1-31.
- Bowonder, B.; P. Sudhakar and D. Wood. (2002) "E-Engineering: Redefining the Boundaries of the Firm", *International Journal of Information Technology and Management*, 1, 1, 32-49.
- Bowonder, B.; J. Racherla; N. Mastakar and S. Krishnan. (2005) "R&D Spending Patterns of Global Firms", *Research-Technology Management*, 48, 5, 51-9.
- Brusconi, S.; A. Prencipe and K. Pavitt. (2002) "Knowledge Specialisation and the Boundaries of the Firm", *Administrative Science Quarterly*, 46, 4, 597-621.
- Bryant, K. (1998) "Evolutionary Innovation Systems: Their Origins and Emergence as a New Economic Paradigm", in K. Bryant and W. Alison (Editors), *A New Economic Paradigm? Innovation-based Evolutionary Systems*. Canberra: Department of Industry, Science and Resources, 53-84.
- Burgelman, R.; M. Maidique and S. Wheelwright. (Editors), *Strategic Management of Technology and Innovation*. Second Edition, New York: McGraw-Hill.
- Cannon, P. (2005) "Why We Do R&D (A Practitioner's Tale)", *Research-Technology Management*, 48, 5, 10-1.
- Cervantes, M. and D. Guellec. (2002) "The Brain Drain: Old Myths, New Realities", *Observer*, No. 230, 40-2.
- Cohen, J. (2000) *Public Policies in the Pharmaceutical Sector: A Case Study of Brazil*. Human Development Department, The World Bank, LCSHD Paper Series No. 54.
- Cohen, W. and S. Klepper. (1996) "A Reprise of Size and R&D", *The Economic Journal*, 106, 925-52.
- Cohen, W. and D. Levinthal. (1989) "Innovation and Learning: The Two Faces of R&D", *The Economic Journal*, 99, 569-96.
- Courvisanos, J. (2005a) "Technological Innovation: Galbraith, the Post Keynesians and a Heterodox Future", *Journal of Post Keynesian Economics*, 28, 1, Fall, 2005, 83-102.
- Courvisanos, J. (2005b) "A Post-Keynesian Innovation Policy for Sustainable Development", *International Journal of Environment, Workplace and Employment*, 1, 2, 187-202.
- Dahlman, C. and J-E. Aubert. (2001) *China and the Knowledge Economy: Seizing the 21st Century*. Washington D.C.: World Bank Institute.
- David, P.; B. Hall and A. Toole. (2000) "Is Public R&D a Complement or Substitute for Private R&D? A Review of the Econometric Evidence", *Research Policy*, 29, 497-529.
- Dolfsma, W. (2006) "IPRs, Technological Development, and Economic

- Development", *Journal of Economic Issues*, 40, 2, 33-42.
- Ettlie, J. (2000) *Managing Technological Innovation*. New York: John Wiley & Sons.
- Ettlie, J. and A. Rubenstein. (1987) "Firm Size and Product Innovation", *Journal of Product Innovation Management*, 4, 2, 89-108.
- Flowers, E. (1998) "The Ethics and Economics of Patenting the Human Genome", *Journal of Business Ethics*, 17 (15) 1737-45.
- Fox, J. (2001) "Why Is It So Difficult to Develop Great Ideas and Inventions in Australia?", in *The Alfred Deakin Lectures: Ideas for the Future of a Civil Society*. Sydney: ABC Books, 228-38.
- Freeman, C. (1994) "The Economics of Technical Change: A Critical Review Article", *Cambridge Journal of Economics*, 18, 4, 463-514.
- Freeman, C. and L. Soete. (1997) *The Economics of Industrial Innovation*. Cambridge, Mass: MIT Press.
- Gansler, J. (1980) *The Defence Industry*. Cambridge, MA: MIT Press.
- Gerschenkron, A. (1962) *Economic Backwardness in Historical Perspective*. Cambridge. MA: Harvard University Press.
- Gilboy, G. (2004) "The Myth Behind China's Miracle", *Foreign Affairs*, 83, 4, July/August. www.foreignaffairs.org/2004/4.html
- Jolly, V.K. (1997) *Commercializing New Technologies: Getting from Mind to Market*. Boston: Harvard Business School Press.
- Kuemmerle, W. (1997) "Building Effective R&D Capabilities Abroad", *Harvard Business Review*, 75, March-April, 61-9.
- Legge, J.M. (2000) "Review of *The Economics of Industrial Innovation*", *Review of Political Economy*, 12, 2, 249-55.
- Legge, J. and K. Hindle. (2004) *Entrepreneurship: Context, Vision and Planning*. Basingstoke: Palgrave Macmillan.
- Lessig, L. (2004) *Free Culture: How Big Media Uses Technology and the Law to Lock Down Culture and Control Creativity*. New York: Penguin.
- Mansfield, E.; J. Rapoport; J. Schnee; S. Wagner and M. Hamburger. (1972) *Research and Innovation in the Modern Corporation*. London: Macmillan.
- Miotti, L. and F. Sachwald. (2003) "Co-operative R&D: Why and With Whom? An Integrated Framework of Analysis", *Research Policy*, 32, 1481-500.
- McDaniel, B. (2002) *Entrepreneurship and Innovation: An Economic Approach*. Armonk, NY: M.E. Sharpe.
- Menke, M. (1997) "Managing R&D for Competitive Advantage", *Research-Technology Management*, 40, 6, 40-2.
- Meyer-Krahmer, F. and G. Reger. (1999) "New Perspectives on the Innovation Strategies of Multinational Enterprises: Lessons for Technology Policy in Europe", *Research Policy*, 28, 7, 751-76.
- Nelson, R. (1992) (Editor) *National Innovation Systems: A Comparative Study*. Oxford: Oxford University Press.
- OECD. (2005) *OECD Science, Technology and Industry Scoreboard 2005*. Paris: Organisation of Economic Cooperation and Development.
- Pavitt, K. (1994) "Key Characteristics of Large Innovating Firms", in Dodgson, M. and Rothwell, R. (Editors), *The Handbook of Industrial Innovation*, Cheltenham: Edward Elgar, 357-66.
- Pavitt, K.; M. Robson and J. Townsend. (1987) "The Size Distribution of Innovating Firms in the UK: 1945-1983",

- Journal of Industrial Economics*, 35, 297-316.
- Rosenberg, N. (1982) *Inside the Black Box: Technology and Economics*. Cambridge, UK: Cambridge University Press.
- Rosenberg, N. (1990) "Why Do Firms Do Basic Research (With Their Own Money)?" *Research Policy*, 19, 165-74.
- Salter, A. and B. Martin. (2001) "The Economics Benefits of Publicly Funded Basic Research: A Critical Review", *Research Policy*, 30, 509-32.
- Scherer, F.M. (1980) *Industrial Market Structure and Economic Performance*. Houghton Mifflin: Boston.
- Schonfeld & Associates. (2005) *R&D Ratios and Budgets*. June 2005 Edition, Riverwoods, Illinois: Schonfeld & Associates, Inc.
- Schumpeter, J. (1934) *The Theory of Economic Development*, Cambridge, Mass: Harvard University Press. [German original 1911].
- Schumpeter, J. (1942) *Capitalism, Socialism and Democracy*. New York: Harper & Row.
- Skea, J. (1994) "Environmental Issues and Innovation", in M. Dodgson and R. Rothwell (Editors), *The Handbook of Industrial Innovation*. Cheltenham: Edward Elgar, 421-31.
- Smith, K. (1998) "Innovation as a Systemic Phenomenon: Rethinking the Role of Policy", in K. Bryant and W. Alison (Editors), *A New Economic Paradigm? Innovation-based Evolutionary Systems*. Canberra: Department of Industry, Science and Resources, 17-51.
- Tsai, K.-H. and Wang, J.-C. (2005) "Does R&D Performance Decline with Firm Size? A Re-examination in Terms of Elasticity", *Research Policy*, 34, 966-76.
- Tassey, G. (1997) *The Economics of R&D Policy*. Westport, CT.: Quorum Books.
- Tidd J.; J. Bessant. and K. Pavitt. (2005) *Managing Innovation*. Third Edition, Chichester: John Wiley and Sons.
- von Hippel, E. (2005) *Democratizing Innovation*. Cambridge, MA: MIT Press.
- Webster, E. (1999) *The Economics of Intangible Investment*. Cheltenham: Edward Elgar.
- White, M. (2005) *Fruits of War: How Military Conflict Accelerates Technology*. London: Simon & Schuster.
- Whiteley, R. (1994) "Cims Industrial Research Institute's Annual R&D Forecasting Trends", *Research-Technology Management*, 37, 1, 22ff.

Websites

- Harvard Business School Working Knowledge. hbswk.hbs.edu/index.jhtml
- World Intellectual Property Organisation. www.wipo.int/portal/index.html.en.
- Intellectual Property Australia. www.ipaustralia.gov.au.
- R&D Magazine. www.rdmag.com/
- Ewing Kauffman Foundation. www.kauffman.org/items.cfm?itemID=678.
- Management of Innovation and New Technology Research Centre. mint.mcmaster.ca/mint/about.htm.
- Journal of the Industrial Research Institute. www.iriinc.org/webiri/Publications/RTM/rmbibl.htm.

Jerry Courvisanos
Centre for Regional Innovation and
Competitiveness, University of Ballarat
Victoria, Australia.
j.courvisanos@ballarat.edu.au

Sexual Harassment in the Workplace

Jérôme Ballet and Françoise de Bry

Introduction

Sexual harassment designates a set of practices of a sexual nature that are exercised in the workplace. Legal definitions are relatively recent. According to Farley (1978), the concept emerged during 1974 in the course of a discussion about women and work at Cornell University. But if the term is recent, the phenomenon is not (ILO 1992; Fitzgerald & Shullman 1993). For example, it is historically well known that domestic service employees have been subjected to sexual abuse on the part of their employer. (Segrave 1994).

Sexual harassment in the workplace encompasses practices that are now condemned, but which were formerly tolerated and even legal (Ballet & de Bry 2007). However, legislation has taken different orientations in various countries. The debate in Europe has been greatly relayed by the media, which have painted an excessively moralistic picture of feminists who have embarked on a “war of the sexes” or an “anti-sex crusade”, in an attempt to counter Victorian standards (Baer 1996; Möller 1999; Hauserman 1999; Saguy 2003). The European Union clearly has a perspective that is slightly different from that of the United States.

The American Perspective

Sexual harassment remains a notion which it is not easy to define and delimit. It is as diverse as workplace fondling vis-à-vis having wall calendars of naked pinup girls. The first constitutes a direct personal injury while the second may only constitute at best an indirect injury. Yet, in the second case, the effects should not be neglected. The women subjected to such posting may feel that their

status is degraded.. Such wall postings suggest that the male employees are preoccupied with sex and that they consider their female colleagues from the point of view of sexual intercourse (Posner 1999).

Two aspects of sexual harassment should be distinguished. On the one hand, there are cases where the persons are the victims of a supervisor who, in return for sexual favours, offers them advantages, or on the contrary, dismisses them or reduces their salary in case of refusal. This type of situation applies to the notion of *quid pro quo* in the American legislation. It corresponds historically to the first cases treated by legislation in the 1960s and 1970s. But this restrictive vision leaves out numerous situations of women who are the victims of sexist jokes or of other conducts which cannot be directly integrated into the notion of *quid pro quo*. Unlike the first ones who had at their disposal means of action at law, the latter could not lay claim to anything (Lipper 1992).

In 1977, in the *Barnes v. Costle* case (561 F.2d 983), a federal circuit court rendered a judgment according to which the remarks and entreaties, when they are linked to risks of dismissal, do constitute sexual harassment. The field of sexual harassment then broadened with the introduction of the notion of *hostile work environment*, which is precisely aimed at taking into account other conducts. However, this broadening also produced vagueness about the delimitation of the concept (Foulis & McCabe 1997).

These aspects impacted on American legislation, which influenced policies elsewhere, including areas as diverse as the UK and India. (Husbands 1992). In the United States, sexual harassment charges are associated with Title VII of the Civil Rights Act of 1964, which bans discriminations at work on the grounds of race, skin colour, sex or national origin. The Equal Employment Opportunity Commission (EEOC) was

created in 1965 with a view to administering and enforcing this act.

US legislation associates sexual harassment with discrimination. MacKinnon (1979), who was one of the attorneys in the *Barnes* case, has published a pioneering book on the subject, which emphasizes this conjunction between sexual harassment and discrimination. More recently, the assimilation of harassment to discrimination has been defended forcefully by Superson (1993). Her main argument is that sexual harassment should be regarded not as an injury to a specific victim, but rather as an attack on the group of all women. This argument raises two problems (Basu 2003). On the one hand, it is not obvious that an attack on a black and lesbian woman should be considered as an attack on women in general, rather than on all the black people or homosexuals. On the other hand, Superson's argument rests on the idea that harassment stems from the feeling that the victim is inferior. It may be convenient to categorize harassment on the basis of the effects on the victim rather than on the basis of the motivations of the harasser.

Problems remain, however (Hajdin 2002; Basu 2003). Four such problems are especially notable (Basu 2003). The first problem is that harassment does not only concern women. Complaints are also being lodged by men; sometimes against women. According to the EEOC, complaints by men in the US represented 9.1 per cent of all complaints in 1992, increasing to 13.7 per cent in 2001. In Europe, according to the European Foundation for the Improvement of Living and Working Conditions, they represented 5 per cent of the complaints in 2000 (EFILWC 2000). Understanding sexual harassment as discrimination based on the gender divide does not seem completely adequate.

Secondly, harassment practices happen, to an appreciable extent, among same-sex individuals. According to a survey carried out by the U.S. Merit Systems Protection Board (USMSPB 1995), 21 per cent of men who say they have been the victims of harassment were primarily harassed by other men. In France, men are attacked in 76 per cent of cases by men, while women are attacked in 40 per cent of cases by women (Leymann 2001). Such situations cannot be covered by the principle of sexual discrimination and the gender divide.

A typical case of this interpretation is the verdict of the U.S. Supreme Court, clarified in 1998 in the case of *Oncale v. Sundowner Offshore Services Inc.* (523 US 75 [1998]). This case involved a man being subject to a hostile work environment from other men. The Supreme Court interpreted Title VII as including sexual harassment among same-sex persons.

Thirdly, in certain instances, harassment is exercised in a vertical way by an employer, without distinction and with the same intensity towards the men and women who are under his command (Paul 1990, Epstein 1985). In that case, harassment cannot qualify as discrimination on the grounds of gender, but has more to do with class.

Fourthly, some persons are attacked not because of their gender but because of their sexual orientation. The case of Mr. Medina Rene, working for the MGM Grand Hotel in Las Vegas is significant (Abelson 2001; Talbot 2002). This openly homosexual man had been the victim of harassment by his male colleagues for several years. As his complaints to the hotel management were ignored he took legal action. The U.S. district court decided that the harassment suffered by Mr. Rene was not based on his gender but on his sexual orientation. Consequently his case could not be covered by Title VII of the Civil Rights Act. However, the U.S. Court of

Appeal of the Ninth Circuit went beyond this judgment and ruled, on 24th September 2002, that the case of *Rene v. MGM Grand Hotel Inc* had to be re-examined.

In view of these different problems, defining harassment as gender discrimination is inadequate as a general view. For this reason, European legislation has taken a different road.

European Perspective

While taking note of the American influence, European legislation has tried to examine harassment in its own way. The clearest definition is that proposed by Article 1 of the European Commission (1991):

“It is recommended that the Member States take action to promote awareness that conduct of a sexual nature, or other conduct based on sex affecting the dignity of women and men at work, including conduct of superiors and colleagues, is unacceptable if:

- (a) such conduct is unwanted, unreasonable and offensive to the recipient;
- (b) a person's rejection of, or submission to, such conduct on the part of employers or workers (including superiors or colleagues) is used explicitly or implicitly as a basis for a decision which affects that person's access to vocational training, access to employment, continued employment, promotion, salary or any other employment decisions; and/or
- (c) such conduct creates an intimidating, hostile or humiliating work environment for the recipient.”

This approach was developed further in the late 1990s (European Commission 1998) and into the twenty-first century (European Parliament et al 2002). For instance, consider the Directive of 23rd September 2002 “*on the implementation of the principle of equal*

treatment for men and women as regards access to employment, vocational training and promotion, and working conditions”. This directive defines sexual harassment in its Article 2 as a situation “*where any form of unwanted verbal, non-verbal or physical conduct of a sexual nature occurs, with the purpose or effect of violating the dignity of a person, in particular when creating an intimidating, hostile, degrading, humiliating or offensive environment*”. That directive provided for 5th October 2005 as the final date for implementation of policy directions by member states. As Table 1 indicates, the 1990s was a key decade for legal action on sexual harassment in the Western world.

Table 1. Laws Against Sexual Harassment and/or Relevant Court Decisions

	Sex equality laws	Sexual harassment laws	Court decision
United States	1964	1980	1987
EC/EU	1976	1990/2002	
Austria	1979	1992	
Belgium	1978	1992	
Denmark	1978		1989
Finland	1986	1995	
France	1983	1992*	
Germany	1980	1994	1986
Ireland	1977	1998	1985
Italy	1991	1996*	
Luxembourg	1981	2000	
Netherlands	1983	1994	
Portugal	1989		
Spain	1980	1989	
Sweden	1979	1991	
UK	1975	1995	1986

Source: Adapted from Zippel (2006) Criminal, Penal Code.

Like the American legislation, the European legislation includes two notions of harassment, *quid pro quo* and *hostile work environment*. However, it does not restrict the area of application by not assimilating it into discrimination legislation. It mainly rests on the notion of the *workers' dignity* while protecting sexual freedom. Sexual harassment is not perceived as a question of inappropriate sexual conduct but is examined from the point of view of marginalizing practices against individuals (Zippel 2006).

The transposition of the directive into the national law of the European Union Member States was carried out very imperfectly on account of the different histories of the national legislations (see Table 1), but mostly on account of the different interpretations made of the conduct involving an attack on dignity in relation to the conducts coming within the space of sexual freedom. For example, the political decision-makers in Spain and in France have explicitly rejected the American conception on the pretext that the Mediterranean culture includes flirting and seduction as forms of normal interactions between men and women in the workplace (Valiente 1998, Saguy 2003, Jenson and Sineau 1995). But they have also interpreted the European directive very restrictively.

For example, the French law only points to “the fact of harassing someone *with the purpose of* obtaining sexual favours”. In the French law, sexual harassment is defined as “the dealings of any person whose intention is to obtain sexual favours for him/her or for a third party” (Labour Code Article L122-46). Such a definition departs from the European directive in at least two ways.. On the one hand, it rules out the notion of *hostile environment* to retain only that of *quid pro quo*. On the other hand, according to such a definition, it is the motivation of the actions that permits delimitation rather than a category of actions. Now, the American and European legislations qualify harassment on the basis of the experience of the persons harassed, and not on the basis of the motivations of the person who perpetrates the acts of harassment (Zippel 2006). Such a perspective, by putting the stress on the dealings of the harasser rather than on the sex of the victim, permits the legislation to go beyond the first two criticisms levelled above against the American legislation.

However, such a choice is not without its ambiguity. Thus, in the French law, an

employee working in a mixed office who displays a calendar with naked women or if there is a situation with repeated sexual overtures may be interpreted as intended acts of harassment if the motivation is explicitly to obtain sexual favours. It could be regarded as insignificant acts if they do not reflect any sexual intention, for even though these acts are certainly uncalled-for, they are not reprehensible. Harassment will actually be considered only on the basis of the following double condition: the acts must be repeated and the complainant must be in a position to prove that they are really intended to obtain sexual favours. Given that it is difficult for the complainant to produce evidence, a great number of complaints for sexual harassment are rejected.

The legislation completes the protection system by defining moral harassment. The latter reflects the *hostile environment* aspect. In the French law, moral harassment is defined as “repeated dealings whose purpose or effect is a deterioration of the working conditions that is likely to undermine the rights and the dignity of the person, to affect his/her physical or mental health, or to compromise his/her professional future” (Labour Code Article L122-49). Moral harassment may therefore be viewed either on the basis of the motivation for the acts or on the basis of the effects, which differentiates it from sexual harassment for which only the motivation was required. Furthermore, given the elements of the definition (dignity, health, professional future), it clearly appears that sexual harassment constitutes a sub-class of moral harassment. That being the case, even if the acts committed are aimed at obtaining sexual favours but do not appear as such at first sight, it will not be possible to qualify them as sexual harassment, but only as moral harassment. We could consider that this overlap is not a real problem, yet the sentences incurred in one case or the other are

quite different – a year in prison and a 15,000-euro fine for sexual harassment, and a year in prison and a 3,750-euro fine or only one of the two sentences for moral harassment (see for example in the French case Articles 152-1 and 222-33 of the Criminal Code). This will incite the persons indicted to have their acts qualified as moral harassment insofar as it is reprimanded less severely than sexual harassment.

The European way, though promising, is still a long way from constituting a homogeneous bloc throughout all the Member Countries. The opening of the E.U.'s internal borders to new Member Countries obviously involves a harmonization process which is very likely to be long.

Extent of the Phenomenon

Sexual harassment mostly affects women at one time or another of their professional life (Schneider et al 1997 Fitzgerald & Ormerod 1993). The fact that there is no data collection at a centralized level, or even the fact that there are absolutely no data for a great number of countries, does not help to assess the extent of the phenomenon. The existing data on sexual harassment only correspond to complaints lodged and not to all the cases of harassment. In its latest publication, the European Foundation for the Improvement of Living and Working Conditions (EFILWC 2005), which based its works on the complaints investigated in Europe, notes that 6 per cent of the employees are affected by sexual harassment. The acts of sexual harassment seem to be increasing, as are all the forms of violence in the workplace. There are significant differences between the various countries (Table 2).

This table underlines the men's share of complaints is not insignificant and seems to be getting closer to that of women in most countries. However, in the countries where the number of cases recorded is the greatest,

the women's share is far higher than the men's. These countries are also those which have the broadest legislation. In other words, the small difference between the women and men in terms of sexual harassment could result, in most cases, from quite a restrictive conception of harassment, which would be limited to established acts, and not to the creation of a hostile work environment. Conversely, the countries which integrate completely the notion of *hostile environment* could see the share of cases against women increasing. In some countries, the environment qualified as hostile would not be taken into account as it would reflect a "culture" or a male "power" which is considered "normal", whereas in the most "severe" countries, the existence of a hostile environment would be a statutory offence.

Table 2. Awareness (or Reports) of Sexual Harassment at Workplace by Country

	Total (%)	Men (%)	Women (%)
EU Total	6	5	8
Finland	12	8	14
Netherlands	11	10	13
Sweden	10	6	13
United Kingdom	8	7	9
Belgium	7	5	9
Greece	7	7	7
Austria	6	4	8
Ireland	6	5	7
Germany	5	3	7
Denmark	4	3	5
Luxembourg	4	5	6
Portugal	3	2	3
France	3	2	4
Spain	2	2	2
Italy	1	2	1

Source: Adapted from EFILWC (2005).

In the United States, the EEOC provides the statistics of the cases which it investigates.. In 2001, 15,475 cases were recorded, 13.7 per cent of which concerned men.

It is necessary to take a certain number of precautions when interpreting the official data on sexual harassment.

Firstly, the national legislations of the various countries are not harmonized and reveal a more or less narrow conception of the problem. The comparisons between the various legislations has given rise to a number of studies (Bernstein 1994; Zippel 2006 for the United States and the European Union; Elman 1996, 2000 for the US and Sweden; Cahill 2001 for the US and Austria; Saguy 2003 for the US and France; Roggeband and Verloo 1999 for Spain and the Netherlands; Mazur 2002 for France and Spain; Baer 1995, Kuhlman 1996 and Zippel 2006 for the US and Germany; also see Aeberhard-Hodges 1996 for a comparison between the different continents). The American legislation, which is closely associated with the notion of *discrimination*, leaves a certain number of cases out. The French legislation qualifies certain conducts as moral harassment rather than as sexual harassment. These acts will therefore appear only rarely within the rubric of sexual harassment. In a good many cases, the complaints are simply rejected. Article 222-33 of the French Criminal Code holds that only the harasser who misuses the authority which his functions confer on him and uses “orders, threats, constraints or serious pressures” to “obtain sexual favours” is penalized. Apart from these scenarios, the acts of violence undergone do not have any legal existence and there is therefore no appeal possible for the victim.

The French legislation on sexual harassment therefore appears to be very restrictive. The complainant could perhaps have won the case if he had lodged a complaint for moral harassment, but it is clear that the sanctions against the employer are different. The American legislation would certainly have permitted to qualify this case on the grounds of discrimination.

Secondly, the very perception which the individuals affected have of the acts

pertaining to sexual harassment may lead to significant differences with the reality of what is legally described as being part of sexual harassment. The growing figures do not necessarily reveal the increase in the number of cases of sexual harassment. Indeed, the number of complaints can increase while the number of cases decreases. The data are then the reflection of these contradictory developments (Basu 2003). Antecol and Cobb-Clark (2002) underline that if, between 1978 and 1994, the cases of sexual harassment recorded by the U.S. Federal Government did not rise significantly, the facts and conducts qualified as sexual harassment by the persons concerned increased considerably. It seems that a structural change in the characterization of sexual harassment took place over the past thirty years so much so that the developments retrace partly the change of perception rather than the real increase in the number of cases.

Thirdly, there are still numerous victims who do not dare to take the plunge of declaration and indictment. In certain cases, there is even a reversal of positioning witnessed between the attacker and the victim; in that case, the latter will blame himself/herself for his/her conduct, his/her dress, etc., and will feel responsible for the acts committed by the attacker whose responsibility he/she will tend to minimize. In some countries, such as France, the reversal of complaints strategy seems to be developing quickly. The complaints for sexual harassment which come to nothing are followed by complaints on the part of the “harasser” for false accusation. Now, false accusation is penalized more severely than sexual harassment, which leads numerous victims of harassment into a difficult situation (AVFT 2004). Such a procedure obviously incites to a reduction in the number of complaints as long as the person harassed is

not convinced that he/she will be in a position to assert his/her legitimate right.

Impacts on the Management of Organisations and Human Resources

The implications of sexual harassment on the companies' human resource management has not been extensively researched. Most research has focused on the legal aspects and the question of harassment measures, but little has been done on the impact of harassment. The potential costs of sexual harassment for companies could be substantial. For example, the U.S. Merit Systems Protection Board estimated these costs at US\$327m between 1992 and 1994 for only the cases of federal agencies, and excluding the legal procedural costs.

A usual method to assess the impact of sexual harassment on the organisations consists of assessing its impact on the persons, on their work satisfaction and on their desire to remain in their employment or to leave. Work satisfaction may actually be regarded as an important indicator of intentions regarding employment (Freeman 1978, Gordon and Denisi 1995, Laband and Lentz 1998). However, the notion of *satisfaction* involves a strong subjectivity. Work satisfaction is also linked to other characteristics concerning the workers (such as age) or to the characteristics determining the status of the employment (such as the size of the establishment, self-employment status, union status) (Heywood & Wei 2001; Shields & Ward 2000).

Research in psychology tend to show that a drop in work satisfaction is correlated with absenteeism (Clegg 1983), with a lower labour productivity (Mangione and Quinn 1975) and a rise in the number of mental and physical health problems.

The links between sexual harassment, work satisfaction and the fact of leaving one's employment are only embryonic. This is

probably due to the fact that the estimation of the impact of sexual harassment is relatively new. Only two surveys establish these links between harassment and employee behaviour - Laband and Lentz (1998) on the women lawyers in the United States and that of Antecol and Cobb-Clark (2001) on the women who have joined the U.S. Armed Forces. These two surveys conclude, on the one hand, that the women victims of sexual harassment are less satisfied in their work than non harassed women and that they express a stronger will to leave their employment.

The survey of the EFILWC (2005) underlines that important differences may be noted according to industry of employment (Table 3).

Table 3. Awareness of Sexual Harassment at Workplace by Industry

	Total (%)	Men (%)	Women (%)
Agriculture	2	2	4
Manufacturing	4	3	5
Electricity	2	1	5
Construction	2	1	5
Trade	5	3	7
Hotels,	13	11	15
Restaurants			
Transportation	7	6	10
Financial	4	4	5
Services			
Business	5	5	5
Activities			
Public Admin.,	6	5	7
Defence			
Health,	9	9	9
Education			

Source: Adapted from EFILWC (2005).

The differences between the various branches of industry suggest several avenues of reflection, which remain to be examined more thoroughly. First of all, the sectors where gender segregation is high could give rise to a more marked harassment of women (building industry, electricity, military). Some studies emphasize that the women working in an environment where there is a majority of men report more cases of sexual harassment in

comparison with the other women (USMSPB 1995; Fitzgerald et al 1997). Nevertheless, Antecol and Cobb-Clark (2001) note that the women working in an environment where there is a majority of women, if they globally express greater work satisfaction, are also more likely to leave their employment if they undergo a form of sexual harassment.

Harassment is not only perpetrated by employees, in sectors where employees are in contact with external customers, the risk of harassment could increase (hotels, restaurants). Thus, in the Netherlands, in 2002, the Netherlands' Work Situation Survey (TAS) indicates that 3.1 per cent of persons have been the victims of sexual harassment on the part of their colleagues and 6.7 per cent have been the victims of sexual harassment on the part of customers. Depending on sex, 11.4 per cent of women and 3.0 per cent of men have been the victims of customers. These results suggest that certain branches of industry, where contact with the customers is central, could record a higher rate of harassment since it is not limited to the relations between colleagues or between employees and supervisors. Moreover, this fact brings to light an important limit of legislation in most countries as the latter precisely confines itself to the relations inside the company and does not take into account the relations with the customers or suppliers.

However, on that point too, the impact of sexual harassment remains to be assessed. Antecol and Cobb-Clark (2001) underline that, in the case of women working in the military, sexual harassment on the part of customers does not seem to affect the women's work satisfaction and does not have an influence on their decision to leave their employment. On its part, the AVFT (1990) has indicated that there are cases of sexual harassment which involve both customers and the superiors. The employers, who do not

want to loose their customers but get them to sign contracts, do not hesitate to exert pressures on their female employees in order to urge them to satisfy all the desires of the customers. The impact of this double harassment—moral harassment on the part of the employers and sexual harassment on the part of the customers, and even double sexual harassment on the part of both the employers and the customers—also requires further assessments. Given the extent of the phenomenon, the sparse results that are currently available call for more extensive sector-based surveys.

Conclusion

The fight against sexual harassment largely goes through company level implementation of measures aimed at reducing its occurrence. The company's ability to create, through its organisation, a climate in which sexual harassment is not tolerated constitutes a decisive factor in reducing the incidence of sexual harassment (Williams et al. 1999). Women who undertake training on harassment are less likely to experience it (Antecol & Cobb-Clark 2001). Without any constraint of the law, it is highly likely that nothing will happen if regulation is left to voluntary codes. One of the few studies on the impact of different laws on the implementation of measures in the companies is that of Zippel (2006) who compared the United States and Germany. The decision of the Supreme Court in the United States in the *Meritor Savings Bank v. Vinson* case in 1986 and several legislative provisions which followed, contributed towards a wave of implementation policies across US where consultants and the lawyers play an important advisory role in the process of policy implementation. The effects of information and consciousness-raising, as well as the effects of fear linked to the legal cost which sexual harassment can represent for the

employer, have greatly contributed towards a preventive approach in the USA. In contrast, in Germany, the law has almost no impact on employer practices.

Selected References

- Abelson, Reed. (2001) "Men, Increasingly, are the Ones Claiming Sexual Harassment by Men", *New York Times*, June, 10, 1.
- Aeberhard-Hodges, Jane A. (1996) "Sexual Harassment in Employment: Recent Judicial and Arbitral Trends", *International Labour Review*, 135, 5, 499-533.
- Antecol, Heather and Deborah A. Cobb-Clark. (2001) The Sexual Harassment of Female Active-Duty Personnel: Effects on Job Satisfaction and Intentions to Remain in the Military, Discussion Paper Series, IZA DP 379, Institute for the Study of Labor, Germany.
- Antecol, Heather and Deborah A. Cobb-Clark. (2002) *The Changing Nature of Unemployment-Related Sexual Harassment: Evidence from the U.S. Federal Government (1978-1994)*. Discussion Paper Series, IZA DP 619, Institute for the Study of Labor, Germany.
- AVFT. (1990) *De l'abus de pouvoir sexuel. Le harcèlement sexuel au travail*. Paris, La découverte.
- AVFT. (1999) *Harcèlement sexuel: Améliorons les lois*. www.AVFT.org
- AVFT. (2004) *Notes sur la situation des femmes poursuivies ou condamnées en dénonciation calomnieuse*. www.AVFT.org
- Baer, Susanne. (1995) *Würde oder Gleichheit? Zur angemessenen grundrechtlichen Konzeption von Recht gegen Diskriminierung am Beispiel Sexueller Belästigung am Arbeitsplatz in der Bundesrepublik Deutschland und den USA*, Baden-Baden, Nomos.
- Baer, Susanne. (1996) „Pornography and Sexual Harassment in the EU“, in *Sexual Politics and the European Union: The New Feminist Challenge*, Amy R. Elman. (ed) 51-66, Providence RI, Berghahn Books.
- Ballet, Jérôme and Françoise de Bry. (2007) *Place des femmes, place de la féminité dans l'entreprise*. Paris, Seuil.
- Bartel, Ann P. (1981) "Race Differences in Job Satisfaction: A Reappraisal", *Journal of Human Resources*, 16, 2, 294-303.
- Basu, Kaushik. (2003) "The Economics and Law of Sexual Harassment in the Workplace", *Journal of Economic Perspectives*, 17, 3, 141-157.
- Bernstein, Anita. (1994) "Law, Culture and Harassment", *University of Pennsylvania Law Review*, 142, 4, 1227-1331.
- Cahill, Mia. (2001) *The Social Construction of Sexual Harassment Law: National and Organizational Effects*. Burlington VT, Ashgate.
- Clark, Andrew E. (1997) "Job Satisfaction in Britain", *British Journal of Industrial Relations*, 34, 2, 189-217.
- Clegge, Chris W. (1983) "Psychology of Employee Lateness, Absence, and Turnover: A Methodological Critique and an Empirical Study", *Journal of Applied Psychology*, 68, 88-101.
- Epstein, Richard. (1985) *Takings: Private Property and the Power of Eminent Domain*. Cambridge Mass., Harvard University Press.
- EFILWC. (2005) *Violence, Bullying and Harassment in the Workplace*, European Foundation for the Improvement of Living and Working Conditions, 2005 Report.
- Elman, Amy R. (1996) *Sexual Subordination and State Intervention: Comparing Sweden and the United States*, Oxford, Berghahn Books.
- Elman, Amy R. (2000) *Sexual Harassment Policy: Sweden in European Context*.

- Paper presented at the Twelfth International Conference of Europeanists, Chicago, March 30-April 1.
- European Commission. (1991) *Commission Recommendation on the Protection of the Dignity of Women and Men at Work*. Brussels: Directorate-General for Employment, Industrial Relations and Social Affairs.
- European Commission. (1998) *Sexual Harassment in the Workplace in the European Union*. Brussels: Directorate-General for Employment, Industrial Relations and Social Affairs.
- European Parliament and the European Council. (2002), *Directive Amending Council Directive 76/207/EEC on the Implementation of the Principle of Equal Treatment for Men and Women as Regards Access to Employment, Vocational Training and Promotion, and Working Conditions*, European Parliament and Council Directive 2002/73/EC, OJ L 269, 5 October, pp.15.
- Farley, Lin. (1978) *Sexual Shakedown: The Sexual Harassment of Women on the Job*, New York, McGraw-Hill.
- Fitzgerald, Louise and Ormerod, Alayne J. (1993) "Breaking Silence: The Sexual Harassment of Women in Academia and the Workplace", in *Psychology of Women: A Handbook of Issues and Theories*, Florence L. Denmark and Michele A. Paludi. (eds) Westport, Connecticut, Greenwood Press.
- Fitzgerald, Louise and Shullman, S.L. (1993) "Sexual Harassment: A Research Analysis and Agenda for the 1990's", *Journal of Vocational Behavior*, 42, 5-27.
- Fitzgerald, Louise; Drasgow, Fritz; Hulin, Charles L.; Gelfand, Michele J.; Magley, Vicki J. (1997) "Antecedents and Consequences of Sexual Harassment in Organizations: A Test of an Integrative Model", *Journal of Applied Psychology*, 82, 4, 578-589.
- Foulis, Danielle and McCabe, Marita. (1997) "Sexual Harassment: Factors Affecting Attitudes and Perceptions", *Sex Roles*, 37, 9/10, 773-798.
- Freeman, Ronald B. (1978) "Job Satisfaction as an Economic Variable", *American Economic Review*, 68, 135-141.
- Gordon, Michale E. and Denisi, Angelo S. (1995) "A Re-Examination of the Relationship Between Union membership and Job Satisfaction", *Industrial and Labor Relations Review*, 48, 2, 222-236.
- Gutek, Barbara A. and Dunwoody, Verna. (1987) Sex Object and Worker: Incompatible Images of Women, Conference Paper, Vol. III, 1986-1987, paper 13, Institute for Social Sciences, University of California, Los Angeles.
- Hajdin, Mane. (2002) *The Law of Sexual Harassment*, London, Associated University Presses.
- Hamermesh, Daniel S. (1977) "Economics Aspects of Job Satisfaction", in *Essays in Labor Market Analysis*, O.E. Ashenfelter and W.E. Oates. (eds) New York, John Wiley.
- Hauserman, Nancy R. (1999) "Comparing Conversations about Sexual Harassment in the United States and Sweden: Print Media Coverage of the Case Against Astra USA", *Wisconsin Women's Law Journal*, 14, 1, 45-68.
- Heywood, John S. and Wei, Xiangdong. (2001) "Performance Pay and Job Satisfaction", unpublished working paper.
- Husbands, Robert. (1992) "Sexual Harassment Law in Employment: An International Perspective", *International Labour Review*, 131, 6, 535-559.
- International Labour Office. (ILO). (1992) *Combating Sexual Harassment at Work, Conditions of Work Digest*, 11, 1, Geneva.

- Jenson, Jane and Sineau, Mariette. (1995) *Mitterand et les Françaises: Un rendez-vous manqué*, Paris, Presses de la Fondation Nationale des Sciences Politiques.
- Kuhlmann, Ellen. (1996) *Gegen die sexuelle Belästigung and Arbeitsplatz. Juristische Praxis und Handlungsperspektiven*, Pfaffenweiler.
- Laband, David N. and Lentz, Bernard F. (1998) "The Effects of Sexual Harassment on Job Satisfaction, Earnings, and Turnover Among Female Lawyers", *Industrial and Labor Relations Review*, 51, 4, 594-607.
- Leymann, Heinz. (2001) *La persécution au travail*, Paris, Seuil.
- Lipper, Nicolle. (1992) « Sexual Harassment in the Workplace : A Comparative Study of Great Britain and the United States », *Comparative Labor Law Journal*, 13, 293-342.
- MacKinnon, Catharine A. (1979) *Sexual Harassment of Working Women: A Case of Sex Discrimination*, New Haven CT, Yale University Press.
- Mangione, T.W. and Quinn, R.P. (1975) "Job Satisfaction, Counterproductive Behaviour and Drug Use at Work", *Journal of Applied Psychology*, 60, 114-116.
- Mazur, Amy G. (2002) *Theorizing Feminist Policy*, Oxford, Oxford University Press.
- Mazur, Amy G. (2003) "Drawing Comparative Lessons from France and Germany", *Review of Policy Research*, 20, 3, 493-523.
- Möller, Simon. (1999) *Sexual Correctness: Die Modernisierung antifeministischer Debatten in den Medien*, Opladen, Leske & Budrich.
- Paul, Ellen F. (1990) "Sexual Harassment as Sex Discrimination: A Defective Paradigm", *Yale Law and Policy Review*, 8, 333-365.
- Posner, Richard A. (1999) "Employment Discrimination: Age Discrimination and Sexual Harassment", *International Review of Law and Economics*, 19, 421-446.
- Roggeband, Conny and Verloo, Mieke. (1999) "Global Sisterhood and Political Change: The Unhappy 'Marriage' of Women's Movements and Nation States", in *Expansion and Fragmentation: Internationalization, Political Change and the Transformation of the Nation State*, Kees van Kersbergen, Robert H. Liesbout and Grahame Lock. (eds) 177-194, Amsterdam, Amsterdam University Press.
- Saguy, Abigail. (2003) *What is Sexual Harassment? From Capitol Hill to the Sorbonne*, Berkeley CA, University of California Press.
- Schneider, Kimberly T., Swan, Suzanne and Fitzgerald, Louise. (1997) "Job-Related and Psychological Effects of Sexual Harassment in the Workplace: Empirical Evidence From Two Organizations", *Journal of Applied Psychology*, 82, 3, 401-415.
- Segrave, Kerry. (1994) *The Sexual Harassment of Women in the Workplace, 1600 to 1993*, Jefferson NC, McFarland.
- Shields, Michael A. and Ward, Melanie E. (2000) "Improving Nurse retention in the British National Health Service: The Impact of Job Satisfaction on Intentions to Quit", *Forschungsinstitut zur Zukunft der Arbeit*. (IZA) Discussion Paper, 118.
- Superson, Anita. (1993) "A Feminist Definition of Sexual Harassment", *Journal of Social Philosophy*, 24, 1, 46-64.
- Talbot, Margaret. (2002) "Men Behaving Badly", *New York Times Magazine*, October, 13, 52f.
- USMSPB. (1995) "Sexual Harassment in the Federal Workplace: Trends, Progress and Continuing Challenges", A report to the President and the Congress of the United

States, Washington D.C., U.S. Merit
Systems Protection Board.

Valiente, Celia. (1998) "Sexual Harassment
in the Workplace: Equality Politics in
Post-authoritarian Spain", in *Politics of
Sexuality: Identity, Gender and
Citizenship*, Mariagrazia Rossilli. (ed) 61-
86, New York: Peter Lang.

Williams, Jill Hunter; Louise Fitzgerald and
Fritz Drasgow. (1999) "The Effects of
Organizational Practices on Sexual
Harassment and Individual Outcomes in
the Military", *Military Psychology*, 11, 3,
303-328.

Zippel, Kathrin S. (2006) *The Politics of
Sexual Harassment*. Cambridge:
Cambridge University Press.

Jérôme Ballet

*Centre for Economics and Ethics of the
Environment and Development*

Universite de Versailles

Saint-Quentin-en-Yvelines

France

jballetfr@yahoo.fr

Françoise de Bry

University of Paris 1,

Pantheon—Sorbonne

Paris, France

Small Business and Entrepreneurship Policy

Rachel Parker

Introduction

There are two key approaches to entrepreneurship, each of which has different implications for small business policy (Danson 2002). The first conceives of entrepreneurship as an economic process and can be traced to the work of Joseph Schumpeter who developed the concept of creative destruction to describe the entrepreneurial process that led to the simultaneous elimination of old industries and activities and the creation of new activities through the commercial application of new ideas. While entrepreneurship as a process of creative destruction might include start up activity amongst small firms, it does not exclusively involve small firms as large firms may contribute to the entrepreneurial process through the generation of new knowledge and by assisting in financing the development of new ideas amongst small firms.

Although innovation occurs in large as well as small firms, the literature on small enterprise innovation draws heavily on Schumpeter's depiction of the central role of the entrepreneur in the process of creative destruction, whereby the economic system is transformed from within and new cycles in economic life emerge in which new industries and markets replace old industries and markets. Schumpeter argued that entrepreneurs drove the process of innovation and that innovation was a stimulus to economic development and involved the development of new products, processes, methods of production or new forms of commercial or financial organisation (Schumpeter 1911). At a time when technological development and structural

economic change are occurring at a rapid pace, small firm innovation is seen to be critically important because empirical evidence, although not undisputed, indicates that SMEs make an important contribution to radical innovations in new industries (Nooteboom 1994).

The second view of entrepreneurship focuses on the individual entrepreneur more than the entrepreneurial process. The entrepreneur is depicted as an owner of small businesses, and is regarded as having particular personal characteristics such as self-reliance, individual initiative and self-motivation. Entrepreneurs are also considered to have a behavioural orientation towards the exploitation of new ideas and opportunities. They are the risk takers who are able to see an opportunity and pursue it commercially despite the uncertainty of rewards. The capacity to plan, manage and lead is also seen to be identifying characteristics of entrepreneurs.

Different small business policy approaches arise from these different perspectives on entrepreneurship. Small business policy approaches that emphasise the process by which new ideas are generated and applied commercially arise from the first and broader view of entrepreneurship. Policies designed to generate a population of risk taking and self-motivated individuals with highly developed management and commercial skills are more in keeping with the second approach, which is focused on the individual entrepreneur rather than the entrepreneurial process.

Importance of Small Business and Entrepreneurship

Increased interest in small business and entrepreneurship emerged with the work of David Birch, who in the 1980s declared that small firms created most new jobs in the United States (Birch 1987). More recently, Audretsch (1995) has drawn attention to the

loss of jobs in large US corporations, particularly in the manufacturing sector, and the job creating role of small companies. Small firms have been portrayed as dynamic agents of change in an evolving economic context. As such, the enhanced policy emphasis on entrepreneurial activities and small business can be linked to broad changes in the structure of economic activity and the organisation of production and work in the advanced economies over the last three decades.

During this period there has been a decline in traditional industries such as manufacturing, which have been dominated by large firms, and a marked growth in social, personal, financial and commercial services, as well as a stabilisation or decline in public employment. The overall effect has been a shift in employment towards industry sectors that have a stronger orientation towards small firms. Flexible and decentralised production has become a reality with the development of new technologies, particularly with the use of computers to aid the production process. There has also been an increase in the role of knowledge and the importance of product and process innovation. A set of strategies for managing the workforce has emerged in response to perceived changes in the structure of industry and the organisation of production including 'reengineering' and 'de-layering', which are thought to reflect the increased need for worker participation and autonomy in the new production regime. These developments have broadly transformed the trajectory of economic development in modern economies away from production regimes associated with stability, control, certainty, continuity and low risk in favour of turbulence, flexibility, uncertainty and rapidly changing and high-risk activities. In this environment, the generation of new ideas is thought to be critical, as is the capacity to test

those high-risk and uncertain ideas in a commercial setting.

Audretsch and Thurik (2001) have represented these changes as a shift from a 'managed' to an 'entrepreneurial' economy. In Audretsch and Thurik's conceptualisation of the entrepreneurial economy, there is an emphasis on turbulence, diversity and heterogeneity, which render small flexible enterprises critical to economic success. In the (former) managed economy, large enterprises were well suited to the environment of stability, continuing and homogeneity. However it is that orientation that renders large firms less competitive in the dynamic entrepreneurial economy in which there is a need for variety in economic activities, firms and the genetic make up and personal experiences of people involved in commercial enterprise. Diversity is necessary to improve the potential for generating new ideas and diversity is increased with a large population of small firms.

A perceived advantage of small firms in the changed environment relates to their internal organisation, which is thought to be relatively decentralised and flexible, allowing them to respond to rapidly changing markets. This perspective is well captured by OECD reports on small and medium-sized enterprises (SMEs) that have argued that SMEs are more flexible and entrepreneurial and have a better capacity to respond to changing consumer demand, to adopt different forms of work organisation and to introduce new technologies, products and processes (OECD 2002).

A Critical Approach

Critical approaches to small business and entrepreneurship highlight the limitations of studies that claim the superiority of small firms as a general class. Some literature has questioned the job generating potential of small firms by showing that small firms are

responsible for both the creation and destruction of a large proportion of all jobs such that their contribution to net job creation is not much higher than their share of employment. There is also evidence that labour conditions, hours and rights may be inferior in small firms as employees in small firms have on average lower levels of unionization, lower wages, less training, higher levels of casualisation and there is evidence that they work longer hours than employees in larger firms ((Davis et al 1996:298-299; Loveman & Sengenberger 1991:23; Parker 2001).

The more critical literature on small business has also questioned the view that small firms are more innovative than large firms. Some research indicates that only a small proportion of small firms undertake formal research and development expenditure and that large firms account for most of a nation's R&D expenditure (Freeman & Soete 1997:228). However, there is evidence to suggest that those small firms that do engage in R&D do so at a high level of intensity and with higher productivity than large firms. The few small firms that do engage in research and development do so at very high levels relative to their employment share or sales. High productivity in small firm innovation is particularly apparent in highly innovative and low capital intensity industries, where small firms' innovative output relative to input appears to be higher than for large firms (Nooteboom 1994:338).

Small firms are generally regarded as making the most significant contribution to radical innovations that involve new technological trajectories and depend on significant changes in a firm or its personnel. In contrast, large firms make the most significant contribution to incremental innovation in medium technology industries such as engineering, machine tools and chemicals. Nooteboom (1994:344) has shown

that large firms are generally more successful in basic innovations, while small firms appear to be more successful at translating those basic innovations into commercial applications. As the life cycle of a product progresses and economies of scale and price competition become increasingly important, the role of large firms in innovation becomes more significant. At the same time, small firms remain important to innovation in niche markets.

It also seems that the relative importance of small and large firms in the process of innovation varies between industries. Small firms appear to play a more significant role in those industries where capital intensity, development costs and barriers to entry for new firms are all low (Freeman & Soete 1997:227-241). The literature on invention and innovation therefore indicates that neither small or large firms are more important – their role and contribution differs according to the industry, the different stages of an industry life cycle and the nature of the innovation.

Some research has shown that there is great diversity amongst the population of small firms, with only a small proportion of small firms making a significant contribution to innovation and employment growth (Loveman & Sengenberger 1991:18-19). Other research has indicated that in English speaking countries, small firm activity might be high in sectors of the economy that are not knowledge intensive and which are relatively low paid (Parker 2001). As such, caution needs to be adopted in relying on comparisons of levels of small business and entrepreneurship which do not distinguish between different types of firms and which do not consider the level of small business activity in knowledge activities. The critical approach to small business and entrepreneurship does not suggest that small business and entrepreneurship are

unimportant. However, the critical tradition is suspicious of approaches that treat small business as a single category of firms and which argue that small business as a general class is superior to large firms in terms of employment and innovation.

Policy Approaches

Governments have become increasingly interested in the level of small business and entrepreneurial activity occurring within their economies. This has had a significant influence on policy development in the sense that small business policy has become a major focus of industry policy initiatives in the OECD countries (Storey and Tether 1998, OECD 2002). There appear to be two major approaches to small business and entrepreneurship policy (Parker 2002). One approach has tended to emphasise the need to promote an overall business environment supportive of small firms and entrepreneurial activities. For example, Stevenson and Lundström (2001:11-32) and Verheul et al (2002:43-51) discuss a range of policy factors impacting on entrepreneurship including macroeconomic variables such as taxation, labour market regulation, social security and income policy; regulatory factors such as establishment legislation, bankruptcy policy, administrative burdens, compliance costs, deregulation and competition policy; and cultural factors such as social and cultural norms that support entrepreneurship.

They suggest that governments wishing to promote small business and entrepreneurship should ensure that administrative requirements for corporations do not render it too difficult or expensive to start a firm and that taxation policy encourages self-employment and does not discourage investment in high growth activities. Further, income policy needs to allow for the accumulation of wealth for reinvestment in new activities and labour market policy must

be sufficiently flexible to allow small firms to hire workers and to dismiss them if ventures fail or strategic directions change. According to this view, the size of government and the welfare state should not be so large as to stifle the culture of individual creativity, self-reliance and initiative that is typically linked to entrepreneurial activity.

Measures that fall within this approach are generally designed to improve the framework conditions for entrepreneurship or the business environment for SMEs. Greater market flexibility achieved through reduced government regulation, combined with enhanced market incentives for entrepreneurial activity are regarded as central to the achievement of small firm competitiveness on the assumption that 'failures in entrepreneurship are attributable to maladjustment to market conditions and to lack of economic incentives' (Martinelli 1994:476). Small business policy problems are conceived in terms of over-regulation of the economy, which is thought to stifle initiative and creativity. As a result, support for SMEs has often involved changes in the broader policy environment, which have been introduced in an effort to improve the economic conditions for business and the rewards for entrepreneurial activity.

This approach is consistent with the view that the Anglo-Saxon economies, with their emphasis on individualism and market competition, provide a more competitive economic environment for entrepreneurial activity in the current context. As it involves a relatively 'hands off' role for government this approach might be termed a passive approach to small business and entrepreneurship policy. The Australian government may be regarded as having adopted this approach because a major objective of small business policy has been to reinstate market incentives for entrepreneurial activity, where they might have been eroded

through taxation or regulation, particularly labour market regulation. A range of regulatory reforms introduced from the 1980s and pursued with vigour since the mid 1990s include competition policy, liberalisation of labour markets, and tax reform. Many of these changes have occurred with a specific emphasis on improving the cost competitiveness of Australian business as high costs are perceived as eroding the rewards for entrepreneurial activity. These policy reforms indicate a strong emphasis on market relations in Australia's approach to small business policy.

The second stream of analysis on small business and entrepreneurship policy does not emphasise the promotion of liberal market conditions as a stimulant for entrepreneurship. It instead encourages the adoption of a proactive approach to promoting knowledge generating activities and the commercialization processes of small firms, which are regarded as critical in the knowledge economy. This approach has focused on specific measures in support of SMEs including direct financial support to SMEs, the provision of advisory services to SMEs, the education and training of PhDs in science and technology and linkages between SMEs and publicly funded research institutions (Storey & Tether 1998). Much of the research falling within this approach has focused on new technology based firms and has therefore taken into account the specific needs of firms engaged in knowledge intensive activities (Heydebreck et al 2000; O'Gorman 2003; Storey & Tether 1998).

Policy measures relating to financing recognise the difficulty that SMEs face in gaining access to finance, particularly in financial systems traditionally oriented towards larger firms. In Germany, the state has encouraged public savings banks and credit co-operatives to form a three tier structure with local branches, closely

affiliated to regional and national institutions, taking responsibility for providing long term finance to SMEs at the local level (Vitols 1997:24). In Sweden, access to finance for SMEs has been addressed through technology bridge foundations and public sponsored venture capital funds such as Industrifonden, which operate at a regional level. OECD countries have introduced a variety of programs designed to improve access to venture capital funds and high-risk loans for SMEs (OECD 1998).

Policy measures in OECD countries have increasingly focused on support for intangible investments such as information and advisory services and training. One such example is the National High Technology Mentor Programme established by the Finnish Academies of Technology in 1993. This programme has relied on the voluntary services of senior managers with significant international experience and expertise in a range of areas including business law, human resource management, finance and technology strategy. These senior managers have mentored managers in participating firms for between 5 and 10 days per year with the aim of promoting the growth of Finnish industry. SMEs have been expected to participate in the program for around 7 to 10 years, although higher technology firms have been able to opt to participate for only 1 to 3 years (Leppänen 1998).

Programs to improve innovation and the utilisation of advanced technologies have been prioritised in many countries. In France, a network of technological development advisers has existed to help SMEs identify their technological needs and to link them with providers of technology. There have also been programs to encourage the recruitment of technological personnel in SMEs to help them identify and implement appropriate technologies for their organisation. A National Research Exploitation Agency

(ANVAR) has sought to encourage innovation in products or processes in SMEs by providing interest free loans, which are refundable in the event of success (OECD 1995:167-75).

A further dimension of policies affecting small business include urban and regional issues involving the regional delivery of programs and services for SMEs including research and development support and financial support. The seven regionally based technology bridge foundations in Sweden are an example of regional initiatives to stimulate the commercialisation of knowledge and improve access of small firms to university research.

Conclusion

The promotion of small business and entrepreneurship has become a key objective of governments seeking to achieve competitiveness in the knowledge economy. A range of policy approaches has emerged in order to address the concerns of small business. At one end are policies that form part of a passive approach, which seek to create an overall business environment in support of small firms based on low costs and market competition. A more active approach has been adopted in some countries and has involved policies designed to improve access to finance, skills and technology and has tended to emphasise technology development and commercialisation amongst small firms. The passive approach sits comfortably with that aspect of the small business and entrepreneurship literature that argues that small business is superior as a general class of firms in the rapidly changing knowledge economy. The second approach involves discriminatory policies which focus on providing support for small business engaged in high technology and knowledge intensive activities and as such is more consistent with critical traditions which are suspicious of the

idea that small business is superior as a general class of firms. The second approach is focused on the process of entrepreneurship while the first approach is concerned with encouraging the individual entrepreneur. As such, different small business and entrepreneurship policies represent different view of entrepreneurship.

Selected References

- Audretsch, D. B. (1995) *Innovation and Industry Evolution*. MIT Press, Cambridge, MA.
- Audretsch, D. and R. Thurik. (2001) "What is New about the New Economy: Sources of Growth in the Managed and Entrepreneurial Economies", *Industrial and Corporate Change*, 10, 1, 25-48.
- Birch, D. (1987) *Job Generation in America: How our Smallest Companies Put the Most People to Work*. New York: Free Press.
- Danson, M. W. (2002) "Entrepreneurship and ICT Industries: Support from Regional and Local Policies", *Regional Studies*, 36, 8, 909-919.
- Davis, S. J., Haltiwanger, J. and S. Schuh. (1996) "Small Business and Job Creation: Dissecting the Myth and Reassessing the Facts", *Small Business Economics*, 8, 4, 297-315.
- Freeman, C. and L. Soete. (1997) *The Economics of Industrial Innovation*. London: Pinter.
- Heydebreck, P. M. Klofsten, and J.C. Maier. (2000) "Innovation Support for New-Technology Based Firms: The Swedish Technopol Approach", *R&D Management*, 30, 1, 89-100.
- Leppänen, R. (1998) *High Tech Company Coaching and Rating: Experience from the Mentor Programme*. Espoo, Finland: Office of the Director Mentor Programme, Innopoli Science Centre.

- Loveman, G. and Sengenberger, W. (1991) "The Reemergence of Small-Scale Production: An International Comparison", *Small Business Economics*, 3, 1, 1-37.
- Martinelli, A. (1994) "Entrepreneurship and Management", in N. J. Smelser and R. Swedberg (Editors), *The Handbook of Economic Sociology*. Princeton: Princeton University Press.
- Nooteboom, B. (1994) "Innovation and Diffusion in Small Firms", *Small Business Economics*, 6, 327-347.
- OECD (2002) *Small and Medium Enterprise Outlook*. Paris: OECD.
- OECD (1995) *Best Practice Policies for Small and Medium Sized Enterprises*. Paris: OECD.
- OECD (1998) *Best Practice Policies for Small and Medium Sized Enterprises*. Paris: OECD.
- O'Gorman, C. (2003) "Stimulating High-Tech Venture Creation", *R&D Management*, 33, 2, 177-188.
- Parker, R. (2002) "Coordination and Competition in Small Business Policy: A Comparative Analysis of Australia and Denmark", *Journal of Economic Issues*, 36, 4, 935-952.
- Parker, R. (2001) "The Myth of the Entrepreneurial Economy", *Work Employment and Society*, 15, 2, 239-254.
- Schumpeter, J.A. (1911) *The Theory of Economic Development: An Inquiry into Profits, Capital, Credit, Interest and the Business Cycle*. New Brunswick: Transaction Books, 1983.
- Stevenson, L. and Lundström, A. (2001) *Patterns and Trends in Entrepreneurship/SME Policy and Practice in Ten Economies*. Stockholm: Swedish Foundation for Small Business Research.
- Storey, D. and B. Tether (1998) "Public Policy Measures to Support New Technology-Based Firms in the European Union", *Research Policy* 26, 1037-1057.
- Verheul, I.; S. Wennekers; D. Audretsch and R. Thurik. (2002) "An Eclectic Theory of Entrepreneurship: Policies, Institutions and Culture", in D. Audretsch, R. Thurik, I. Verheulan and S. Wennekers (Editors), *Entrepreneurship: Determinants and Policy in a European-US Comparison*. London: Kluwer Academic Publishers.
- Vitols, S. (1997) "German Industrial Policy: An Overview", *Industry and Innovation*, 4, 1, 15-36.

Rachel Parker

Brisbane Graduate School of Business
Queensland University of Technology

Brisbane, Queensland

Australia

r.parker@qut.edu.au

Stakeholders

Aleksandar Sevic

Introduction

The broadest definition of a stakeholder underscores the existence of an individual or a group who can affect or is affected by the achievement of an organization's objectives (Freeman 1984:46). This definition does not specify requirements necessary for the differentiation between the levels of importance that stakeholders have for the management strategic evaluation of the business environment. Generally applied, it means that all groups or individuals who maintain any sort of relationship with a business entity can be labelled as stakeholders.

In order to find a more comprehensive clarification it has been suggested to examine potential candidates from the following viewpoints: a) the power of the stakeholder to influence the firm; b) the level of her legitimacy vis-à-vis the firm; and c) the urgency of the stakeholder's claim with respect to the firm (Mitchell *et al.* 1995). This approach has emphasized the change in the relevance that stakeholder groups have on the firm at any point in time. It is also comprehensive enough to include marginalized categories that have distant or very specific claims such as the subjects of philanthropic donations or demonstrators.

The early list of stakeholders included shareholders, customers, employees, suppliers, lenders and society. The number of stakeholders' types depends on the approach that researchers apply, which can lead to the inclusion of terrorists, vegetation, nameless sea creatures and unborn generations, among many others (Sternberg 1999). In contrast, Wood (1991) explicitly makes a difference between stakeholder management and

environmental assessment as the separate processes of corporate social responsiveness.

In the 1990s the discussion about what represents stakeholders and how to measure their impact on the performance of an enterprise has become particularly intense. Also, with a strong public appeal of *social responsibility* and the advocating of the firm's awareness regarding the existence of various claimants the voices of critiques have become loud. Is it justifiable for a management of the firm to neglect shareholders to the benefit of stakeholders who cannot be easily identified? How can we measure the impact of stakeholders and who are the principal stakeholders whose interest managers are obliged to uphold? In addition, we should briefly examine the evolution of the acceptance of stakeholders' significance.

Acceptance of Stakeholder Provisions

The Delaware Supreme Court in the *Unocal Corp vs. Mesa Petroleum* case accepted two standards for reviewing takeover defences. In the first instance it was important for directors to provide evidence that there was a threat to corporate policy and effectiveness, while the second pillar assumed that the directors took adequate measures in proportion to the imposed threat. In particular, the board had to simultaneously take account of other constituencies such as creditors, customers and employees (Bainbridge 1992), but their interests are still secondary to those of shareholders (Reynolds 2001).

As the main breakthrough of stakeholder theory in the legislative area authors cite the number of US states that passed non-shareholder constituency (also known as non-monetary factor) statute that serves as amendment to the statutory statement of the director's duty of due care. Directors do not only pay adequate attention to shareholders but also to their non-shareholder constituency

such as employees, consumers, suppliers, creditors and the community where the firm operates. As of the beginning of 1990s more than 25 firms adopted this legal requirement (Bainbridge 1992), having soared up to 38 in mid 1990s (Sternberg 1999).

The Companies Act 1985 (UK) addresses the duty of care by directors to the company, which also underscores the necessity to meet the requirements imposed by employees. This rather general statement does not explicitly clarify whether employees or any other stakeholders would be treated as equals when confronted with shareholders.

The Corporations Act of Australia does not include employee-specific provisions and directors do not have any specific duties to employees. In an empirical study it was found that 74 percent of Australian directors rank shareholders ahead of employees in their duty of care (Francis 1997).

Shishido (1999) claims that the empowerment of shareholders to replace managers in Japan provides them with a significant advantage over other stakeholders such as employees, creditors and trading partners. Therefore, company law can be deemed as focused on shareholders. The differentiation between inside and outside shareholders allows for the conversion of stakeholding rights into shareholding ones. Banks and trading partners in the same industrial groups may have stable (inside) cross-shareholdings, and accordingly, directly impact the decision-making process.

Regardless of the historical, cultural and legislative diversities across the globe it is important to sort out stakeholder groups, apart from shareholders, which exert major influences on and strongly interact with a business entity.

Selected Constituencies

There is not a clear guideline about the nature and purpose of stakeholder theory. According

to Donaldson and Preston (1995) there are three approaches in stakeholder theory. Firstly, a descriptive approach has been used to elaborate the nature of the firm, the stance of managers towards managing, the stance of board members vis-à-vis constituencies and managerial patterns in some corporations. Secondly, instrumental stakeholder theory links theory to descriptive/empirical data in an attempt to provide liaisons between stakeholder management and profitability. And thirdly, the normative approach analyses the function of the firm generously exploiting moral and philosophical doctrines. When contrasting these three aspects of stakeholder theory, the authors stipulate that the central core of the theory is normative, which makes any shareholder based assumption rather morally unsustainable. Following the pivotal normative perception, management believes that all stakeholders have intrinsic value (Donaldson & Preston 1995).

Management

Carol (1979) has introduced the Corporate Social Performance model that includes three dimensions: a) social responsibility categories, b) philosophy of social responsiveness; and c) social issues involved, in an attempt to integrate economics aspects into the social performance program. In this manner, management will be able to analyse the company's position and adequately address stakeholders' claims.

The agency theory has a narrower scope and claims that principals (shareholders and/or bondholders) may have different goals when compared with agents (managers) (Jensen & Meckling 1976). La Porta *et al.* (2002) note that in large corporations it is rather more difficult to expropriate due to public scrutiny, reputation-building, foreign shareholders, and listings on international exchanges. Klapper and Love (2002), using the assumptions rooted in the free cash flow

theory purported by Jensen (1986) claim that big firms with large free cash flows face greater agency problems. In this particular case, management perking may be reduced by increasing dividend payouts if there is a lack of investment projects with positive Net Present Value (NPV). In the public policy context multiple principals, i.e. politicians, cannot expect regulatory agencies to fully comply with principals' requirements, because regulators' actions must reflect electoral goals and are intrinsically unobservable (Spiller 1990). In the similar manner, multiple principals may be represented especially in the board of directors of companies operating in highly regulated industries in order to monitor and achieve specific group's interests (Pfeffer 1972).

Stewardship theory, in contrast, postulates that companies would benefit from the unity of management discretion and control (Donaldson and Davis 1991), because managers are not only driven by personal financial goals, but also by the "the need for achievement and recognition, the intrinsic satisfaction of successful performance, respect for authority and work ethic" (Muth and Donaldson 1998:6) Directors or boards can also avoid using power to certain extent by the introduction of self-control in order to achieve a specific results and act as a good "steward". Anglo legislative systems seem to require that directors' duties be defined in accordance with stewardship theory, because directors are to be trusted in conducting their fiduciary duty towards shareholders (Turnbull 1997). The concept of self-control may be strengthened by the impact of strong organisation ethic deeply embedded, for instance, in Japanese companies.

The Board of Directors may be one of the mechanisms to control for management perking. Nevertheless, not all directors will protect stakeholding rights in the similar way.

Wang and Dewhirst (1992) assume that outside directors should more fervently protect stakeholders' interest, but at the same time, if co-opted by incumbent managers the alignment of interests will lead to rather similar attitudes towards residual and other claimants. Nevertheless, managers and companies must be responsive to social issues in order to be regarded as socially responsible (Clarkson 1995). Moreover, they are not solely responsible to shareholders, because other stakeholders may be also exposed to risk (Kaufman 2002).

Ford and McLaughlin (1984) compare the attitudes of Business School Deans and CEOs towards social responsibility. They conclude that the CEOs demonstrate slightly higher optimism about the corporate acceptance of social responsibility. This may be influenced by the conviction that CEOs do possess corporate clout and realize the importance of the fair treatment of stakeholders. In addition, similar conclusions on what should be priorities in a five-year period reflect the general consensus on social responsibilities between groups that are actively involved in business and those who behave as occasional analysts. Posner and Schmidt (1984) run a survey of 6000 supervisory, middle and executive managers in the US and conclude that they put more emphasis on organisational effectiveness, rather than profit maximization and rank customers and employees ahead of owners as the more important stakeholder groups. Wang and Dewhirst (1992) interview the members of the boards of directors and compare the stakeholder orientation of CEOs vs. non-CEOs directors as well as board insiders vs. outsiders. They conclude that customers and governments are ranked higher with respect to stockholders, employees and society, which can be explained by the main interest in satisfying customers' preferences and compliance with government regulation. Non-CEO directors have stronger stakeholder

orientation when compared to CEOs, while insider directors demonstrate weaker employee orientation than outsiders. The significance of government requests and compliance with local legislation does not indicate that multinational companies have to remain indifferent to the outspoken violation of business ethics and human rights. In 1977 twelve major US companies accepted six principles put forward by Reverend Leon Sullivan, which were expected to eliminate segregation and salary differences, improve general working conditions for non-white employees, promote non-white supervisors and managers, and ensure equal and fair employment practices for white and non-white employees in their affiliates in the Republic of South Africa. US multinationals that had accepted these principles were the forerunners of the widespread US public condemnation of apartheid, which led to the enactment of the Comprehensive Anti-Apartheid Act in 1986 (Post 2002).

Codes of Corporate Conduct, i.e. written statements of principle or policy intended to serve as the expression of a commitment to particular enterprise conduct" (Diller 1999, p. 102), which promote socially responsible behaviour and fair treatment of major stakeholders may appear as firstly, single corporation codes routinely posted on a company's website. Secondly, industry codes are promoted especially for businesses that may have had adverse impact on stakeholders due to the characteristics of the product. Chemical industry in the US accepted a set of ten principles in the Responsible Care as a response to the blemished reputation of the industry after the gas leak at the Union Carbide plant in India that killed more than 4000 people. Thirdly, codes may be created by external groups, such as Social Accountability (SA) 8000 addressing standards at workplaces or CERES principles related to environmental issues (Steiner and

Steiner 2006). Finally, government or multilateral agency codes may be also enacted. UN Global Compact elaborates four areas: human rights, labour, environment and anti-corruption, while OECD Guidelines for Multinational Enterprises also include disclosure, consumer interests, science and technology, competition and taxation (OECD 2000).

Since 1991 CEOs from 35 countries globally address stakeholder issues through the World Business Council for Sustainable Development located in Geneva, Switzerland. WBCSD's Regional Network includes 50 independent organizations which allows for a widespread discussion on economic, environmental and social management in industry. At the same time it offers access to leading multinational companies and sustainable development platforms, with a diversity of projects' focus ranging from environmental protection, market sustainability, labour mobility to corporate social responsibility (WBCSD 2001).

Employees

As active members of the company the interest of employees can be routinely in conflict or in continuity with those of managers or shareholders.

For instance, the English Court of Chancery (1883) in the *Hutton vs. West Cork Rly Company* case confirmed that liberal dealing with employees would provide benefits to the company and, accordingly, ex gratia benefits to employees may not be in conflict with shareholders' interests. Almost 80 years latter this principal has not been applied in *Parke vs. Daily News Ltd.* (Raynolds 2001).

On the other hand, at the beginning of the twentieth century was the *Dodge vs. Ford Motor Company* case. Here profit retaining and production expansion accompanied by price reduction were deemed by plaintiffs as

an altruism towards customers and employees. Even though these actions may be observed as exploiting market opportunities that will provide capital gain for residual claimants at a later stage, the court passed a decision that these actions were unacceptable because the management's primary role is to maximize profits to shareholders. The distributions to non-stockholders were deemed not to be justified. However, incremental improvements in the working condition by some leading US companies may have been greater than what was externally mandated by the government. For instance, Procter & Gamble introduced disability and retirement pensions in 1915, the eight-hour day in 1918, while the work was guaranteed for 48 weeks per annum (Economist 2002). By contrast, a 5-dollar wage introduced by Henry Ford in 1914 may have been aimed at weakening the clout of unionisation rather than improving labour condition.

The specifics of the employees' indivisible human capital produce inherent conflicts (Aoki 1984). It may be reflected by work experience (newcomers vs. close to retirement cohorts) and qualifications (white vs. blue collars). In such an environment it is very difficult to discuss a unique case for employer stakeholding. Following the acceptance of labour contracts employees may be entitled to residual compensations schemes such as occasional renegotiation or the linking of earnings to level profits or improvements in productivity. The third form would be payments in company stock where workers become de facto shareholders in their own companies. Unlike previous residual compensation schemes, the stock payments involve risk that outweighs that faced by shareholders, because bankruptcy would not only result in job loss, but also the effacement of stock holdings (Kaufman 2002).

Apart from employee's terms of trade that are delineated in contracts there are implicit claims that are not mentioned in any agreement, but still expected from stakeholders. These include things such as a clean and safe workplace, future prospects, and job security (Bowen *et al* 1995). Kinder, Lydenberg, Domini and Company Inc is a social investment firm that evaluates the performance of Standard and Poors' 500 and Domini 400 Social Index companies by addressing various stockholder issues on a three-point scale. A company is strong when it establishes favourable union relations, meaningful profit sharing and pension plans, and open communication channels with employees.

Consumers

Donaldson and Preston (1995) claim that in the classical theory an organisation has received resources from investors, employees and suppliers to provide goods and services to customers. In order to be able to constantly produce more effectively and efficiently it is important to foster research and development activities. Sougiannis (1994) claims that an increase in R&D leads to a noticeable increase in profits and an even stronger impact on the companies' valuation. When measuring the after-tax investment value of R&D the author concludes that earnings contain more information on the R&D impact on the price than do the R&D numbers per se. Accordingly, market forces are said to adequately evaluate innovative endeavours.

Another perspective focuses on the effect of product recall on the financial wellbeing of affected firms. The direct loss due to the loss in sales of particular products represents just an initial aspect of consumers' decision making. An additional detrimental impact would be the depletion of a company's goodwill. The goodwill loss is substantial not only for false advertising presentation

(Peltzman 1981) but also for dysfunctional auto parts or potentially harmful drugs (Jarrell and Peltzman 1985). Jarrell and Peltzman conclude that product failure has a broad impact on industry, which is contradictory to the general belief that competing firms would be unequivocally better off after a company's product failure. Hoffer *et al.* (1988) analyse this issue and find little evidence that markets penalise the owners of the firm recalling products or those of its competitors, by reducing the market value of shares.

The significance of customer satisfaction has been incorporated in the derivation of the balanced scorecard advocated by Kaplan and Norton (1992), which supplements classic financial performance measurements with additional non-financial measures such as internal business processes, and learning and growth. The balanced scorecard is expected to provide answers to "four basic questions: a) how do customers see us? (customer perspective); b) what must we excel at? (internal perspective); c) can we continue to improve and create wealth? (innovation and learning perspective) and d) how do we look to shareholders? (financial perspective)" (Kaplan & Norton 1992:72). For each one of these areas management is expected to set specific goals and appropriate measures to gauge the level of their achievement. For instance, from the customer perspective it may be necessary to set up goals for time, quality, performance and services.

A company will strive to provide new innovative products, become a responsive as well as preferred supplier, and maintain constructive partnership with customers reflected, for instance, by a number of co-operative engineering efforts. Companies that used the balanced scorecard as the motivating factor leading to better performance were more successful than enterprises that applied this new measurement tool as a summary of management's visions and accumulated

knowledge (Kaplan and Norton 1993). In order to facilitate the most effective application of the balanced scorecard in the firm the top-down strategy must be implemented. The specialised team will articulate the vision of the company and communicate it to mid-level managers who, in turn, will focus on the setting up of divisions' balanced scorecards. The primary feedback will allow the CEO and the top-level executive team to examine each business unit strategy and communicate to the entire company a clear-cut and revised firm's business strategy. Monthly, quarterly, and annual reviews will be conducted to measure performance and allow for business units to clearly delineate their goals and for employees to link objectives to the company's balanced scorecard (Kaplan and Norton 1996).

Unlike classical viewpoint on the characteristics of modern enterprises, the stakeholder theory assumes that shareholders, employees, customers, suppliers and *other stakeholders* contribute and retrieve benefits from a company (Turnbull 1997). This also indicates that apart from regulatory bodies, communities in which companies operate represent a very important stakeholder.

Community

The relationship between businesses and community is evaluated through the donations to charities, voluntary employees' involvement in communal activities, educational and housing initiatives for the poor, contribution to local economic development, protection of human rights and provision of companies' premises to interested parties (Sen & Bhattacharya 2001; Zairi & Peters 2002; Bennett 2002).

Donations to charities are generally tax deductible, and this strong tax incentive may be the driving force for such an action. However, Webb (1996) failed to establish a

positive relationship between tax rates and gifts. In addition, the collection of complete data sets in similar studies is rather challenging because bigger companies respond to questionnaires more frequently than smaller firms. This also indicates that companies with non-existent or minor donations refuse to respond to questionnaires.

Companies apply a specific marketing strategy in selling luxury products by linking them to charities. Hedonic pleasure and guilt increase the drive for altruistic behaviour, which allows for the application of a new charitable sale approach that reduces the sense of guilt with a simultaneous support for the hedonic pleasurable experience. The usefulness of the product's functionality has been reinforced by the customer's knowledge that someone in need will be supported (Strahilevitz and Myers 1998).

While these strategies maximize profits there are also competing claims that management shirking and pressure by "society" will lead to increased charity donations. However, Navarro (1998) finds that donations are a form of positive advertising. Moreover, firms with lower leverage and larger dividend payouts are leading donors to charity, while no link has been found between management compensation/control and contributions.

In war-torn countries multinational companies (MNC) are expected to assist in the post-conflict reconstruction and the rebuilding of destroyed infrastructure (Bennett 2002). In underdeveloped countries MNC may provide drugs free of charge to eradicate debilitating diseases, such as the river blindness by Merck or the distribution of anti-AIDS drugs by Coca-Cola in Africa (Russo & Wertheimer 2004; Simms 2002). While companies do not realise any short-term pecuniary benefits, the association of brand and/or company with corporate social responsibility provides long-term benefits.

Individual donors seek recognition for their donations, which serves as a confirmation of their status and a signal to other peer groups. This may occasionally lead to rivalry between benefactors (Economist 2006). Private provision of public goods would reduce the donations of less wealthy individuals searching for status, while more affluent ones will continue giving contributions (Glazer & Konrad 1996).

In order to enhance morale, loyalty and job satisfaction of employees, companies provide a sponsored volunteer program (Graff 2004). It is also believed that volunteers simultaneously acquire employable skills in the contemporary business environment, which further contributes to the development of labour force and an increase in productivity (Burnes & Gonyea 2005). Communities, in turn, profit from enriched community life, an increase in human resources, improvement in services and historical, artistic as well as cultural enrichment of citizens (Graff 2004). In addition, companies can also provide donations to organisations for which employees worked. In the US, more than 80% of large companies have employee volunteer programs, with an ever increasing trend to employ former employees, i.e. retirees, to accommodate a change in the population pyramid and readiness to secure additional revenues after retirement (Burnes & Gonyea 2005).

Further Discussion on Stakeholder Theory

With the separation of ownership and control in MNCs along with a large base of international shareholders the social responsibility of the large capitalist company has improved. Shumpeter believes that the bureaucratisation and team-project approaches to R&D lead to the routinisation of innovation and the ultimate demise of the capitalist firm (Manne 1966). If the

foundation of competitive race is not price competition, but an ever growing drive for inventing new products, technology and setting up new organizational units, as claimed by Shumpeter, in the globalised economy these forces have become even stronger. MNCs set up research centres across several continents and outsource knowledge from various parts of the globe, which makes them very competitive vis-à-vis domestic companies that lack international exposure (Shapiro 2003). Even though research endeavours have been routinised, there is not an apparent lack in innovation (Manne 1966).

However, the inclination in large companies to become more stakeholder oriented can lead to multitasking and profit reduction. Jensen (2002) purports the idea that multiple objectives mean no objectives, because some of them may be contradictory such as the profit maximization and maximum market share. In his opinion, the maximization of the total firm value, i.e. the summation of the equity, debt and any other contingent claims outstanding on the firm, renders society better off. Accordingly, in order to mitigate differences between the shareholder and stakeholder theories the author proposes two theoretical approaches: a) *enlightened value maximization*. It is based on the scorecard of the organization that allows for the evaluation of the success. Nonetheless, there is no clear indication how the goal will be achieved; b) *enlightened stakeholder theory*. It is expected that the goal of maximizing total long-term firm market value is added to the main theoretical framework.

The critique about stakeholder theory may be reflected in the opinion that the number of potential stakeholders is infinite, and therefore it is rather difficult if not impossible to calculate what claims they may have vis-à-vis the company. Moreover, even if stakeholders can be identified it is quite

challenging to identify and measure benefits that will accrue to them. How should we balance interests and conflicts? Unless these conflicting issues are resolved stakeholder theory may obtain support among managers and various business groups that would benefit from diverting some business funds for personal and other socially unjustified causes. Therefore, stakeholding theory undermines the foundation of private property, appears rather value destructive and hinders the smoothing of the principal-agent relationship (Sternberg 1997). In a constructive response Vinten (2001) stresses that the balanced stakeholder theory does take account of de facto owners because it would be against the law to act otherwise. Also, with a passage of time “the company” has changed its position and role in the society. It has become more exposed to corporate raiders, but also decisively more involved in the community.

The remnants of conservatism associated with natural rights of private property as opposed to the working class engulfed in the mechanical process (Veblen 1955), can be relegated to the early XX century business mentality. The complexity of business environment already demands the set-up of board of directors’ committees analysing social responsibility or ethics policies (Steiner and Steiner 2006). In his futuristic vision Tricker (2000) believes that stakeholder theory will evolve into stakeholder concept whereby stakeholder committees will be able to monitor the company’s decision-making process without any voting power to reverse it. They will exercise power through negotiation skills or publicizing issues of concern.

Selected References

Aoki, Masahiko. (1984) *The Co-Operative Game Theory of the Firm*. Oxford University Press, New York.

- Bainbridge, M. Stephen. (1992) "Interpreting Nonshareholder Constituency Statutes", *Pepperdine Law Review*, 19, 971-1025.
- Bennett, Juliette. (2002) "Multinational Corporation, Social Responsibility and Conflict", *Journal of International Affairs*, 55, 2, 393-410.
- Bowen, M. Robert, Larry DuCharme, and D. Shores. (1995) "Stakeholders' Implicit Claims and Accounting Method Choice", *Journal of Accounting and Economics*, 20, 3, 255-295.
- Burnes, Kathy and Judith G. Gonyea. (2005) *Expanding the Boundaries of Corporate Volunteerism: Retirees as a Valuable Resource*. Mobilizing Retirees for Civic Engagement Project. The Center for Corporate Citizenship at Boston College, Boston.
- Carroll, B. Archie. (1979) "A Three-Dimensional Conceptual Model of Corporate Performance", *The Academy of Management Review*, 4, 4, 497-505.
- Clarkson, B. E. Max. (1995) "A Stakeholder Framework for Analyzing and Evaluating Corporate Social Performance", *Academy of Management Review*, 20, 1, 92-117.
- Diller, Janelle. (1999) "A Social Conscience in the Global Marketplace? Labour Dimensions of Codes of Conduct, Social Labelling and Investor Initiatives", *International Labour Review*, 138, 2, 99-129.
- Donaldson, Lex and James H. Davis. (1991) "Stewardship Theory and Agency Theory: CEO Governance and Shareholder Returns", *Australian Journal of Management*, 16, Number, 1, 40-64.
- Donaldson, Thomas and Lee E. Preston. (1995) "The Stakeholder Theory of the Corporation: Concepts, Evidence, and Implications", *Academy of Management Review*, 20, 1, 65-91.
- Economist. (2002) *Special Report: Lots of it about—Corporate Social Responsibility*, London, December 14.
- Economist. (2006) *One Up for LVMH*. London, October 7.
- Ford, Robert, and Frank McLaughlin. (1984) "Perceptions of Socially Responsible Activities and Attitudes: A Comparison of Business School Deans and Corporate Chief Executives", *Academy of Management Journal*, 27, 3, 666-674.
- Francis, Ivor. (1997) *Future Direction – The Power of the Competitive Board*. FT Pitman, South Melbourne.
- Freeman, Edward R. (1984) *Strategic Management: A Stakeholder Approach*. Pitman, Boston.
- Glazer, Amihai, and Kai A. Konrad. (1996) "A Signaling Explanation for Charity", *American Economic Review*, 86, 4, 1019-1028.
- Graff, Linda. (2004) *Making a Business Case: for Employer-Supported Volunteerism*. Volunteer Canada, Ontario.
- Hoffer, E. George, Stephen W. Pruitt, and Robert J. Reilly. (1988) "The Impact of Product Recalls on the Wealth of Sellers: A Reexamination", *Journal of Political Economy*, 96, 3, 663-670.
- Jarrell, Gregg, and Sam Peltzman. (1985) "The Impact of Product Recalls on the Wealth of Sellers", *Journal of Political Economy*, 93, 3, 512-536.
- Jensen, Michael C. (1986) "Agency Costs of Free Cash Flows, Corporate Finance and Takeovers", *American Economic Review*, 76, 2, 323-329.
- Jensen, M.C. (2002) "Value Maximization, Stakeholder Theory, and the Corporate Objective Function", *Business Ethics Quarterly*, 12, 2, 235-256.
- Jensen, M.C. and William H. Meckling. (1976) "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership

- Structure”, *Journal of Financial Economics*, 3, 4, 305-360.
- Kaplan, S. Robert, and David P. Norton. (1992) “The Balanced Scorecard – Measures that Drive Performance”, *Harvard Business Review*, 70, 1, 71-79.
- Kaplan, S. Robert, and David P. Norton. (1993) “Putting the Balanced Scorecard to Work”, *Harvard Business Review*, 71, 5, 134-142.
- Kaplan, S. Robert, and David P. Norton. (1996) “Using the Balanced Scorecard as a Strategic Management System”, *Harvard Business Review*, 74, 1, 75-85.
- Kaufman, Allen. (2002) “Managers’ Double Fiduciary Duty: to Stakeholders and to Freedom”, *Business Ethics Quarterly*, 12, 2, 189-214.
- Klapper, F. Leora, and Inessa Love. (2002) “Corporate Governance, Investor Protection, and Performance in Emerging Markets”, *World Bank Working Paper* No. 2818.
- La Porta, Rafael, Florencio Lopez-de-Silanes, Andrei Shleifer, and Robert W. Vishny. (2002) “Investor Protection and Corporate Valuation”, *Journal of Finance*, 57, 3, 1147-1170.
- Manne, Henry G. (1966) *Insider Trading and the Stock Market*, The Free Press, New York.
- Mitchell, Ronald K.; Bradley R. Agle; and Donna J. Wood. (1997) “Toward a Theory of Shareholder Identification and Salience: Defining the Principle of Who and What Really Counts”, *Academy of Management Review*, 22, 4, 853-886.
- Muth, M. Melinda and Lex Donaldson. (1998) “Stewardship Theory and Board Structure: A Contingency Approach”, *Corporate Governance: An International Review*, 6, 1, 5-28.
- Navarro, Peter. (1988) “Why Do Corporation Give to Charity?”, *Journal of Business*, 61, 1, 65-93.
- OECD. (2000) *The OECD Guidelines for Multinational Enterprises*, DAFNE/ IME/ WPG 9.
- Peltzman, Sam. (1981) “The effects of FTC Advertising Regulation”, *Journal of Law and Economics*, 24, 403-448.
- Pfeffer, Jeffrey. (1972) “Size and Composition of Corporate Boards of Directors: The Organization and Its Environment”, *Administrative Science Quarterly*, 17, 2, 218-228.
- Posner, Z. Barry, and Warren H. Schmidt. (1984) “Values and the American Manager: An Update”, *California Management Review*, 26, 3, 202-216.
- Post, E. James. (2002) “The ‘Iron Law’ of Business Responsibility Revisited: Lessons from South Africa”, *Business Ethics Quarterly*, 12, 1, 265-276.
- Reynolds, Adam. (2001) “Do ESOPS Strengthen Employee Stakeholder Interests?” *Bond Law Review*, 31, 1, 95-108.
- Russo, B. James, and Albert I. Wertheimer. (2004). *Pharmaceutical Executive*, 24, 4, 40-48.
- Sen, Sankar, and C.B. Bhattacharya. (2001) “Does Doing Good Always Lead to Doing Better? Consumer Reactions to Corporate Social Responsibility”, *Journal of Marketing Research*, 38, 2, 225-243.
- Shapiro, Alan C. (2003) *Multinational Financial Management*, John Wiley and Sons Inc, Hoboken NJ.
- Shishido. (1999) “Japanese Corporate Governance: The Hidden Problems of the Corporate Law”, *Berkeley Olin Program in Law & Economics*, Working Paper Series No. 23.
- Simms, Jane. (2002) “Business: Corporate Social Responsibility – You Know it Makes Sense”, *Accountancy*, 129, 1311, p. 48.

- Sougiannis, Theodore. (1994) "The Accounting Based Valuation of Corporate R&D", *Accounting Review*, 69, 1, 44-68.
- Spiller, T. Pablo. (1990) "Politicians, Interest Groups, and Regulators: A Multiple-Principals Agency Theory of Regulation, or 'Let Them Be Bribed'", *Journal of Law and Economics*, 33, 1, 65-101.
- Steiner, A. George, and John F. Steiner. (2006) *Business, Government, and Society: A Managerial Perspective, Text and Cases*. McGraw-Hill Irwin, Boston.
- Sternberg, Elaine. (1997) "The Defects of Stakeholder Theory", *Corporate Governance. An International Review*, 5, 1, 3-10.
- Sternberg, Elaine. (1999) *The Stakeholder Concept: A Mistaken Doctrine*, Foundation for Business Responsibilities.
- Strahilevitz, Michal, and John G. Myers. (1998) "Donations to Charity as Purchase Incentives: How Well They Work May Depend on What You Are Trying to Sell", *Journal of Consumer Research*, 24, 4, 434-446.
- Tricker, Bob. (2000) "Editorial 2020 Vision: On Corporate Structures in 21st Century", *Corporate Governance. An International Review*, 8, 1, 2-6.
- Turnbull, Shann. (1997) "Corporate Governance: Its Scope, Concerns and Theories", *Corporate Governance: An International Review*, 5, 4, 180-205.
- Veblen, Thorstein. (1958) *The Theory of Business Enterprise*. Mentor Books, New York.
- Vinten, Gerald. (2001) "Shareholder versus Stakeholder—Is there a Governance Dilemma?", *Corporate Governance. An International Review*, 9, 1, 36-47.
- Wang, Jia, and Dudley H. Dewhurst. (1992) "Board of Directors and Stakeholder Orientation", *Journal of Business Ethics*, 11, 2, 115-123.
- WBCSD. (2001) *Stakeholder Dialogue: The WBCSD's Approach to Engagement*, Geneva, Switzerland.
- Webb, J. Natalie. (1996) "Corporate Profit and Social Responsibility: 'Subsidization' of Corporate Income under Charitable Giving Tax Laws", *Journal of Economics and Business*, 48, 4, 401-421.
- Wood, J. Donna. (1991) "Corporate Social Performance Revisited", *Academy of Management Review*, 16, 4, 691-718.
- Zairi, Mohamed and John Peters. (2002) "The Impact of Social Responsibility on Business Performance", *Managerial Auditing Journal*, 17, 4, 174-178.

Websites

KLD. www.kld.com

OECD. www.oecd.org.

United Nations Global Compact. www.unglobalcompact.org.

World Business Council for Sustainable Development. www.wbcsd.org

Aleksandar Sevic
School of Business
Trinity College Dublin
Dublin, Republic of Ireland
a.sevic@tcd.ie

Tax Evasion and Tax Avoidance

Margit Schratzenstaller

Introduction

Tax evasion includes all illegal attempts to conceal taxable activities or objects from tax authorities, tax avoidance all legal activities to reduce the tax liability. Both phenomena can be subsumed under the term "tax aversion". As a consequence of growing international integration and openness of financial markets, tax aversion on an international scale (tax flight) is increasingly gaining in importance: The liberalization of international capital transfers offers manifold opportunities for international tax arbitrage by shifting mobile income earning activities or incomes abroad.

Methods of Tax Evasion and Avoidance

Within a given jurisdiction taxes are evaded either by under- or not reporting the tax base (e.g. interest income earned by private households) to tax authorities, or by shifting taxable activities (e.g. labor or goods trade) into the informal sector. Income and property taxes can be legally avoided by reducing investment and labor supply respectively; consumption tax avoidance requires a reduction of consumption or a substitution of taxed by non-taxed goods.

It can be assumed, however, that international tax flight is of higher quantitative importance than domestic tax aversion. Tax flight is closely interconnected with international tax competition and the existence of offshore financial centres (OFC; also referred to as "tax havens"). Tax havens are defined as politically stable territories where no or hardly any banking and financial regulations exist. Often they are dependent territories of industrialized countries. Some tax havens generally apply low tax rates on personal and corporate income and property,

whereas others grant favorable tax conditions to non-residents only. Generally they have strict secrecy rules and refuse to co-operate with foreign tax administrations; and the disclosure of any information about the activities, in some cases even about the identity of foreign investors is denied (OECD 1987).

Also a large number of "onshore tax havens" with regular tax systems but no or relatively low source taxes for capital income realized by foreign investors, in some cases combined with tight banking secrecy laws (e.g. Switzerland or Luxembourg), qualify as host countries for cross-border portfolio investors seeking to evade income taxation in their residence countries. After in 1984 the United States abolished the source tax on saving income accruing to non-residents, many countries followed suit and now levy no or only a very low source tax on non-residents' interest income to increase their attractiveness for foreign financial investment. Supported by insufficient information exchange and international co-operation between national financial authorities, an increasing number of individual taxpayers shift private savings abroad to evade taxation in their home countries.

Legal tax flight undertaken by transnational corporations which is facilitated by international "fair" or "regular" tax competition and "unfair" or "harmful" tax competition is much more multi-faceted.

In fair tax competition countries are underbidding each others' regular corporate tax rates, which leads to a long-term decrease of corporate tax rates in all countries involved. From 1996 to 2003, the average corporate tax rate decreased from 39% to 32% in the European Union, from 38% to 31% in the OECD (KPMG 2003). A similar trend can be observed in the developing countries where the average corporate tax rate

dropped from 30% to 20% during the past decade. Transnational corporations make use of international tax differentials to reduce the total corporate tax bill by shifting profits from high-tax to low-tax or no-tax countries. Profit-shifting strategies are rather complex and intransparent in practice, but they all rely on transfer pricing and thin capitalization as the basic devices (Hines and Rice 1994): Subsidiaries located in high-tax countries are charged excessive transfer prices (prices for the intra-firm supply of goods and services) or are supplied with loans by subsidiaries located in low-tax countries. Thus, profits in high-tax locations are lowered artificially, whereas they are increased in low-tax locations of a multinational enterprise.

Unfair tax competition according to a widely-used definition of the OECD (OECD 1998) is characterized by the following features: Tax privileges are exclusively granted to foreign investors, and no real economic activity is required. Tax rules are intransparent; in the extreme case they can be negotiated individually. Moreover, there is no or only insufficient information exchange and co-operation with foreign tax authorities. Two groups of tax havens offering unfair tax regimes to multinational companies can be distinguished:

1. Countries with regular company tax systems which grant special preferential tax regimes to transnational corporations (“onshore tax havens”). Transnational corporations are allowed to establish subsidiaries which carry out special activities or services for the parent company and are taxed at special low tax rates: E.g. a financial holding which administers a transnational corporation’s capital assets or a subsidiary which acts as an insurance company (so-called captive) insuring specific risks for the transnational corporation (OECD 2000). These preferential tax regimes are offered by industrialized countries as well as by transition

or developing countries. One prominent example are the so-called coordination centres in Belgium which are engaged in financial transactions for the parent company which is located in a high-tax country; the profits transferred to these coordination centres are taxed at an extremely low tax rate. At the end of the 1990’s, about 350 coordination centres resided in Belgium. Many transition and developing countries have established special economic zones in which foreign investors enjoy special tax exemptions.

2. Offshore tax havens offer the legal opportunity to establish subsidiaries (offshore or base companies) to which profits can be transferred from high-tax countries (OECD 1987). An increasing number of major transnational corporations use offshore companies to reduce their overall tax burden (Kay & King 1990). Most of these tax havens are small territories located in the immediate neighborhood of large industrialized countries. U.S. business can use 41 countries and regions as tax havens, e.g. Barbados or the Cayman Islands (Hines and Rice 1994). Crucial is the treaty network established between tax havens and major industrial countries. These bilateral double taxation treaties reduce or eliminate withholding taxes in the tax haven, so that profits shifted to base companies are subject to extremely favorable tax conditions. By re-routing profits via countries included in the treaty network of a tax haven, transnational corporations can minimize the tax burden on profits which are finally repatriated to the parent company located in a high-tax country.

Tax Aversion Theory and Empirical Evidence

Essentially two approaches to explain tax aversion economically have evolved in the relevant literature. The “tax wedge”- or “tax burden”-approach contends that tax aversion

is fostered by high differentials between before- and after-tax income. According to this explanation, the incentives to shift income earning activities into the shadow economy, to conceal tax bases or to change behavioural patterns to avoid taxation increase with the tax burden (Enste & Schneider 2000). The well-known “Laffer-theorem” is based on a similar notion: It holds that an increase of the tax rate above a certain maximum rate cannot yield additional tax revenues. On the contrary, tax rate increases necessarily lead to decreasing overall tax revenues because firms and households intensify tax avoiding or tax evading efforts. The validity of the Laffer-theorem can be questioned, however: Whether private households decrease their labor supply as a consequence of an increase in income taxes depends on the size of the income effect compared to the substitution effect. A tax-induced decrease of the labor supply only occurs if the substitution effect dominates, i.e. if working hours are substituted by leisure, as neoclassical theory predicts. An income effect on the other hand can cause households to increase labor supply, as net income reductions caused by higher income taxes must be compensated by an increase of working hours.

The neoclassical theory of tax evasion (Allingham & Sandmo 1972) holds that the extent of tax evasion is contingent on the existing system of control and sanctions which determines expected costs and benefits of tax evasion. The more intense monitoring and the higher the penalties for tax evaders are the more effectively tax subjects are deterred from tax evasion.

Both views – the tax-wedge view and the neoclassical theory of tax evasion – are complementary explanations for tax aversion. From an economic perspective tax aversion can be viewed as the result of the net gains which are determined by the associated costs

(the penalty if caught or the cost of behavioral changes) and benefits (savings on tax payments). Therefore tax compliance depends on the probability of being detected, the level of penalties, the marginal tax rate, the relative size of the tax base and the costs of behavioral adjustments.

Over the last decade it has been increasingly acknowledged in the literature that the extent of tax aversion cannot be explained solely on the basis of pure economic arguments. Important non-economic determinants of tax aversion are the perceived fairness of the whole tax system as well as citizens’ tax morale, which in turn is influenced by political and socio-cultural factors, like the general attitude towards the state or social norms (Pommerehne et al 1994). Moreover the purposes for which tax money is used and the degree to which tax payers are involved in decision-making concerning public spending play a role. Finally, also a country’s political system and development stage must be considered because they determine its revenue collecting and tax enforcement potential.

Empirical work done to underpin the existing theoretical explanations for tax aversion has been mainly undertaken with respect to tax evasion in the USA and yields mixed results. According to an econometric study by Clotfelter (1983), the level of income as well as marginal tax rates are positively correlated with tax evasion, a result that cannot be confirmed by Slemrod (1985). Witte and Woodbury (1985) find that the probability of detection and the level of penalties have a significant positive effect on tax compliance, whereas Dubin/Wilde (1988) do not find clear-cut evidence on the significance of the probability of detection. According to surveys conducted by Alm et al (1992) tax compliance improves if tax payers feel that they at least partly benefit themselves from public expenditures financed

by tax revenues. Experiments carried out by Bosco and Mittone (1997) show that the individual propensity to evade taxes is restrained by moral constraints if tax yields are spent for redistributive purposes. Experimental studies could also confirm cross-country differences in tax morale and tax compliance (Alm/Sánchez/de Juan 1995). A survey of private firms in several Eastern European countries finds that there is a significant correlation between the extent of corruption in the public sector and tax evasion (Johnson et al 2000).

It is even more difficult to find empirical evidence on the actual relevance of tax avoidance. Several empirical studies conducted in the last 30 years (e.g., Lall 1973; Grubert & Mutti 1991) confirm the use of profit shifting strategies by transnational corporations; the total amount of taxes avoided is unknown, however. There is no conclusive empirical evidence in the sensitivity of investment and labor supply to variations in the level of taxation. Regarding the reaction of labor supply to (tax-induced) changes in net income, empirical results hint at a rather inferior role of income taxes (e.g. Moffitt & Wilhelm 1998). Particularly male labor supply seems to be quite tax-insensitive because the income effect clearly dominates the substitution effect (e.g. Hamermesh & Rees 1993).

The actual amount of taxes evaded or avoided is highly debated and—due to the nature of the problem—must rely on rather crude estimations. According to the Taxpayer Compliance Measurement Program developed by the US-American Internal Revenue Service 40% of US-households underpaid federal income taxes in 1988 (Andreoni et al 1998). It is estimated that private wealth held in tax havens reached one third of global GDP at the end of the 1990's (Oxfam 2000).

Consequences

of Tax Evasion and Tax Avoidance

The consequences of tax aversion are discussed rather controversially in the literature. “Unfair” tax regimes and the opportunities they present to avoid taxes have come to be perceived unanimously as a problem by economists and governments as they distort competition and capital flows between countries. “Fair” tax competition, however, often is assessed as beneficial because it forces governments to cut tax rates and consequently public expenditures. The same differentiation frequently is made between tax avoidance and tax evasion. Unlike tax evasion, tax avoidance is viewed as a legitimate attempt to escape a tax burden which is perceived as “excessive”. This perspective is based on the perception of the government as “Leviathan” (Brennan & Buchanan 1980) who exploits tax subjects by maximizing tax revenues to serve the selfish interests of politicians and bureaucrats. From this view, restrictions for allegedly wasteful national governments to collect taxes are desirable. A minority of authors, however, who view governments as altruistic and benevolent dictators trying to maximize citizens’ utility argue that the negative consequences of tax aversion outweigh possible positive effects.

According to these authors the evasion of capital income taxes by private households violates social equity. Whereas all other income—particularly labor income—is taxed within the progressive personal income tax, capital income remains tax-free or is subject to low source-taxes only. An increasing number of countries (e.g. the Scandinavian countries) have been introducing dual income tax systems since the beginning of the 1990's: Capital income is taxed at a low flat tax rate to decrease the incentive to move capital abroad and to evade capital taxation (Cnossen 1999). These dual income tax systems are

connected with equity problems as well, however.

Moreover, tax aversion negatively impacts on government budgets because it erodes states' revenue potential in the long run. Governments may be forced to react by cutting transfer payments and/or public investment expenditures. Thus tax aversion can lead to a sub-optimal provision of public goods and services for firms and private households (e.g. Zodrow & Mieszkowski 1986). Negative distributional consequences arise if transfer payments are cut for poor households whereas rich households benefit from lower capital taxes. If revenue losses due to tax aversion are compensated by shifting the tax burden to less mobile tax bases which cannot escape taxation (private consumption and labor), the redistribution potential of tax systems is curtailed, and the tax burden is increasingly unevenly distributed among socio-economic groups (Sinn 1994).

If exit options allowing to escape taxation are only available for certain socio-economic groups respectively for certain income-earning activities, or if only certain pressure groups succeed in influencing tax provisions in their own interest, the general tax morale and the overall acceptance of tax systems are undermined. Finally, elaborate tax evasion schemes increase the administrative costs of tax collection.

Transnational corporations which reduce their total tax liabilities by shifting profits to low-tax or no-tax countries are privileged against enterprises whose activities are limited to their home country and that do not have equal tax-reducing options. Thus not only the equity principle of taxation is violated, but transnational corporations also gain a competitive advantage over domestic enterprises. Moreover transnational corporations act as free-riders who profit from the public infrastructure provided by

high-tax countries, but do not contribute their fair share to its financing.

Alluring private incomes or corporate profits by means of favorable tax regimes can be interpreted as a specific type of a "beggar-my-neighbour-policy": It distorts the allocation of overall taxable income and therefore capital tax revenues across countries and violates inter-nation equity.

Underdeveloped and transition countries are put under even more severe fiscal strain by tax aversion than industrialized countries. Their revenue raising capacities are limited because of an often poor administrative infrastructure and the lack of effective monitoring and control possibilities. Also corruption in the public sector restricts revenue collection in many developing and transition countries (World Bank 1997).

Thus these countries are not able to tax a large portion of domestic private households' capital income and firms' profits respectively. As tax subjects cannot be effectively hindered in transferring taxable income and portfolio investment abroad, developing countries, according to current estimations, lose 50 billion US-\$ annually in tax revenues (Oxfam 2000). As shown above, the fierce competition for foreign investment within the group of developing countries has depressed the average corporate tax rate considerably below that of developed countries. Tax competition has also resulted in the establishment of special economic zones in these countries where generous tax exemptions are granted or even considerable subsidies are paid to foreign investors. Thus domestic enterprises are disadvantaged compared to multinationals which hampers the development of the domestic enterprise sector. Also, multinational companies use the public infrastructure without contributing to public budgets. The resulting revenue losses curtail the ability of underdeveloped countries

to invest in public infrastructure and social services.

Governance Policies and Perspectives

According to the prevailing view that international tax competition and tax avoidance are desirable in principle, regulation on the national and the supranational level currently is mainly concerned with fighting tax evasion.

Many industrial countries apply specific tax collection methods for special kinds of incomes to prevent tax evasion. One example is the collection of income taxes on wages by employers, another example are source taxes for domestic interest income collected the banks (e.g. in most EU member states). An increasing number of countries levy final withholding taxes on domestic interest incomes at a comparatively low level to decrease the incentive for international tax flight (dual income taxes). To secure full declaration of interest income some countries have established a system of automatic information that oblige banks to provide information to fiscal authorities on the identity of domestic recipients of interest income and on the amount of interest income received (e.g. Australia, Canada, the USA or the UK; OECD 2000a). Other governments, however, refuse to adopt such information systems in defence of strict banking secrecy laws (e.g. Austria or Luxembourg).

Tax fraud is considered a criminal act in most countries. Therefore tax evaders are threatened with penalties (fines or imprisonment). The level of penalties normally depends on the amount of taxes evaded, on the frequency of tax evading actions and on the methods applied (i.e. the “criminal energy” involved). The intensity of audits and the level of sanctions vary considerably among industrial countries but generally are considered relatively low (Alm et al 1995). A more effective prevention of

tax evasion therefore requires the intensification of controls to detect e.g. moonlighting or hidden capital income and heavier sanctions. In light of the more recent empirical research results on the determinants of tax evasion sketched above penalties must be supplemented by structural reforms providing for more justice within tax systems: e.g., by equally taxing all types of incomes and by distributing the overall tax burden more evenly among socio-economic groups. The implementation of dual income tax systems appears to be counterproductive with respect to overall tax compliance and tax morale from this perspective.

On the supranational level several initiatives to reduce the extent of tax aversion have been started since the end of the 1990's. They are part of the growing efforts to cope with the potential dangers concomitant to a liberalized global financial system all supranational institutions are increasingly concerned about. Concerning tax evasion on an international scale, there is a growing insight that effective countermeasures require an increasing degree of international cooperation. Currently, international automatic information exchange about interest income by non-residents is stipulated only by a few countries within double taxation treaties (OECD 2000a). Most OECD-countries do not take part in any international automatic exchange of information on bank interest payments, however. Therefore several initiatives are underway on the supranational level since the end of the 1990's. After lengthy negotiations the member states of the EU agreed on a directive on the effective taxation of interest income in 2003. According to this multilateral agreement most of the member states will take part in a system of automatic international information exchange on the interest incomes earned by non-residents from 2005 on.

The US-American IRS forced banks in important European host countries (e.g. Switzerland and Luxembourg) to supply information about the identity and the interest income of US-citizens from 2001 on by threatening them with the loss of their licenses for the US-American financial market. This measure is particularly interesting as it is not bilateral in a classical sense (where two governments are involved), but rather constitutes an agreement made between one branch of the private sector of one country and an official governmental institution of another.

Another approach that can be labelled as “naming and shaming” was chosen by the OECD towards tax havens which are non-OECD-members (OECD 1998, 2000). A black list was compiled including non-cooperative tax havens, initially with the goal to coerce them into abandoning harmful tax regimes. Due to the influence of the USA who refuse to limit national tax sovereignty, the initiative is now aiming at making the targeted tax havens to commit to transparency and to provide information on request. Although the number of non-cooperative tax havens has decreased significantly by now, the effectiveness of this initiative is questionable, as no change of tax practices is required.

The EU has been targeting tax arbitrage by transnational companies which are using unfair tax practices in EU member states since 1997 (European Commission 1997). A code of conduct was adopted by all member states to roll back harmful tax practices and to prevent the introduction of new ones (standstill). During the last years some progress could be achieved. However, the effectiveness of the code is limited as it is only soft law and therefore cannot be enforced with sanctions. Also it does not cover the tax havens associated with the EU member states. Concerning unfair preferential

tax regimes in OECD member states the OECD is proceeding in a similar way (OECD 1998 and 2000), relying on informal and non-binding coordination: It has identified a number of tax practices which are to be abolished by the end of 2005. A regulation of “fair” tax competition within regular tax systems, however, is not on the agenda any more. The latest move in this direction made by the so-called Ruding Committee to introduce minimum corporate tax rates in the EU (Ruding Committee 1992) was strongly opposed by the member states.

To sum up, it must be stated that the problem of international tax flight is not tackled adequately by the policies currently pursued on the international level. Information rights and monitoring capacities of states must be strengthened, for example by loosening banking secrecy regulations. International cooperation and mutual assistance in tax questions as well as international information exchange about income and profits realized abroad must be intensified, and tax systems must be adjusted to the changing international environment.

Selected References

- Allingham, M.G. and Agnar Sandmo. (1972) “Income Tax Evasion – A Theoretical Analysis”, *Journal of Public Economics*, 1, 3-4, 323-338.
- Alm, James; Gary H. McClelland and William D. Schulze. (1992) “Why do People Pay Taxes?”, *Journal of Public Economics*, 48, 1, 21-38.
- Alm, James; Isabel Sánchez and Ana de Juan. (1995) “Economic and Non-economic Factors in Tax Compliance”, *Kyklos*, 48, Fasc. 1, 3-18.
- Andreoni, James; Brian Erard and Jonathan Feinstein. (1998) “Tax Compliance”, *Journal of Economic Literature*, 36, 2, 818-860.

- Bosco, Luigi and Luigi Mittone. (1997) "Tax Evasion and Moral Constraints", *Kyklos*, 50, Fasc. 3, 297-324.
- Brennan, Geoffrey and James M. Buchanan. (1980) *The Power to Tax. Analytical Foundations of a Fiscal Constitution*. Cambridge, MA: Cambridge University Press.
- Clotfelter, Charles T. (1983) "Tax Evasion and Tax Rates: An Analysis of Individual Returns", *Review of Economics and Statistics*, 65, 363-373.
- Cnossen, Sijbren. (1999) "Taxing Capital Income in the Nordic Countries: A Model for the European Union?", *FinanzArchiv*, 56, 1, 18-50.
- Dubin, Jeffrey A. and Louis L. Wilde. (1988) "An Empirical Analysis of Federal Income Tax Auditing and Compliance", *National Tax Journal*, 41, 1, 61-74.
- Enste, Dominik H. and Friedrich Schneider. (2000) "Shadow Economies: Sizes, Causes, and Consequences", *Journal of Economic Literature*, Vol. 38, 1, 77-114.
- European Commission. (1997) *A Package to Tackle Harmful Tax Competition in the European Union*, COM (97) 564 final, Brussels: EC.
- Grubert, Harry and John Mutti. (1991) "Taxes, Tariffs and Transfer Pricing in Multinational Corporate Decision Making", *Review of Economics and Statistics*, 73, 2, 285-293.
- Hamermesh, Daniel and Albert Rees. (1993) *The Economics of Work and Pay*. Fifth Edition. New York: Harper/Collins.
- Hines, James R. and Eric M. Rice. (1994) "Fiscal Paradise: Foreign Tax Havens and American Business", *The Quarterly Journal of Economics*, 109, 1, 149-182.
- Johnson, Simon, et al. (2000) "Why Do Firms Hide? Bribes and Unofficial Activity After Communism", *Journal of Public Economics*, 76, 495-520.
- Kay, John A. and Mervyn A. King. (1990) *The British Tax System*. Fifth Edition. New York: Oxford University Press.
- KPMG. (2003) *Corporate tax rates survey – January 2003*. Amsterdam: KPMG International Tax Centre.
- Lall, Sanjaya. (1973) "Transfer-Pricing by Multinational Manufacturing Firms", *Oxford Bulletin of Economics and Statistics*, 35, 3, 173-195.
- Moffitt, Robert A. and Mark Wilhelm. (1998) "Taxation and the Labor Supply: Decisions of the Affluent", NBER Working Paper 6621, Cambridge MA.: NBER.
- OECD. (2000) *Towards Global Tax Co-operation*, Paris: Organization for Economic Cooperation and Development.
- OECD. (2000a) *Improving Access to Bank Information for Tax Purposes*. Paris: Organization for Economic Cooperation and Development.
- OECD. (1998) *Harmful Tax Competition: An Emerging Global Issue*. Paris: Organization for Economic Cooperation and Development.
- OECD. (1987) *International Tax Avoidance and Tax Evasions*. Paris: Organization for Economic Cooperation and Development.
- Oxfam. (2000) "Tax Havens. Releasing the Hidden Billions for Poverty Eradication", *Oxfam Policy Papers* 6-00, Oxford.
- Pommerehne, Werner; Albert Hart and Bruno S. Frey. (1994) "Tax Morale, Tax Evasion and the Choice of Policy Instruments in Different Political Systems", *Public Finance*, 49, 52-69.
- Ruding Committee. (1992) *Report of the Committee of Independent Experts on Company Taxation*, Brussels/Luxembourg: EC.
- Sinn, Hans-Werner. (1994) "How much Europe? Subsidiarity, Centralization and Fiscal Competition", *Scottish Journal of Political Economy*, 41, 85-107.

- Slemrod, Joel. (1985) “An Empirical Test for Tax Evasion”, *Review of Economics and Statistics*, 67, 2, 232-238.
- Witte, Ann D. and Diane F. Woodbury. (1985) “The Effects of Tax Laws and Tax Administration on Tax Compliance: The Case of the United States Individual Income Tax”, *National Tax Journal*, 38, 1-13.
- World Bank. (1997) *Helping Countries Combat Corruption. The Role of the World Bank*. Washington DC: World Bank.
- Zodrow, George R. and Peter Mieszkowski. (1986) “Pigou, Tiebout, Property Taxation, and the Underprovision of Local Public Goods”, *Journal of Urban Economics*, 19, 3, 356 – 370.

Websites

OECD. www.oecd.org

European Union.

www.europa.eu.int/comm/taxation_customs

Margit Schratzenstaller
Austrian Institute for Economic Research
Vienna, Austria
Margit.Schratzenstaller@wifo.ac.at

Urban and Regional Policy Issues

Oren M. Levin-Waldman

Introduction

Urban and regional policy issues encompass a broad range of issues confronting cities and their larger metropolitan areas, ranging from social and economic to matters of immigration and governance. Though issues vary from city to city, what appears to be common among many throughout the world is the need to attract outside investment as a means of generating economic growth and continued development.

Broadly speaking, the political economy of cities is marked by two principal things. The first is the need to generate growth as the principal vehicle for addressing all other problems plaguing them. The second thing is competition for investment as the vehicle for generating growth. Moreover, this political economy is the inevitable byproduct of the unique political and legal circumstances of most cities—that they are not sovereign entities in their own right, but at best autonomous. The typical urban environment is thus subject to the vicissitudes of its respective broader economic, political, cultural, and social context.

Unlike sovereign nation states, cities do not control their borders, which means that they cannot control who comes and leaves. They don't coin money and they don't conduct foreign policy. In the United States, cities are literally creatures of the states, which means that the states that created them, whether it was through amendment to the state constitution, legislative statute, or the granting of a home rule charter, can destroy them. In European countries, cities are part of further administrative subdivisions of their larger provincial administrative divisions. Consequently, they are autonomous. And yet, what happens within a city often has an

impact on its larger region. This paper will explore the basic problems confronting most cities by situating them within the context of larger urban theory.

Urban Evolution

Urban development has by and large tended to parallel economic developments. The development of many cities parallels the development of Philadelphia, as depicted by Sam Bass Warner (1987). Urban history specifically revolved around a process of private industrialization, which he referred to as privatism. In preindustrial periods, cities were small centers of commerce with relatively homogeneous populations and minimal governmental services. Cities essentially grew through the process of agglomeration—the logical process by which related industries located near one another, and for the purpose of deriving benefit from being in close proximity to one another. It was through the process of urbanization that firms gained access to large and diversified work forces. Douglas Rae (2003) refers to this process as “urbanism” to describe the first general features of cities as they emerged between the 1870s and 1920s.

These features are four-fold. The first is industrial convergence, which entailed the creation of large outward flows of products beyond the cities and their respective regions. It was this export of goods which formed the basis of a steady stream of wages and investment that would energize the city. The second is a diverse fabric of enterprise. The third is a centralized clustering of housing which was to concentrate families of all classes and ethnicities in relatively compact central cities. While the fourth is a myriad of diverse civic organizations outside the business sector that was able to provide yet another layer of cohesion and governance. Urbanism also entailed an important pattern of political integration, which was made

possible by the concentration of leaders from business and civic organizations inside cities on a more or less full-time basis. Although urbanism would lead to a new urbanism characterized by different features steeped in new reality, the substance of the pattern of political integration would nonetheless remain the same, connecting past urban reality to current urban reality.

Urban form has by and large been driven by this evolution. Much of that form was specifically related to industrialization, or what some might refer to as the Fordist model of production, paralleling Henry Ford's assembly line production model. Bruce London and William Flanagan (1976) note that the bulk of literature on spatial ecology of cities—in comparative urban ecology—has been concerned with cities in developing countries. In the “classical school”, Eugene Burgess attempted to answer whether or not cities, despite variations, have an underlying “ideal typical” form which best describes them. In the classical model, which Burgess used to describe Chicago in the 1920s, the city is divided into concentric rings, with the first being the central business district and the fifth being the outer boundaries where the upper classes lived. In between, there was manufacturing activity and working-class living. The poor, it was assumed lived closer to the CBD and Zone 2 where manufacturing occurred because of the high cost of transportation. William Alonso (1983) essentially characterized the composition as following the lines of the bid-rent curve. As land prices dropped and the further one moved away from the CBD, it was likely that the upper classes, because they could afford the higher transportation costs, would move further out so that they could acquire more living space at reduced costs. The poor and the working classes, had no other choice but to live in close proximity to their places of employment. But because they were forced to

live closer to the CBD where land costs were higher, they were in effect forced to live in densely populated areas. London and Flanagan (1976) note that empirical studies of both American and foreign cities have demonstrated that the Burgess construct was both time-bound and culturally specific, and perhaps limited to the descriptions of only a limited selection of North American cities at a particular point in their development.

Similarly, many of the problems plaguing urban areas today are the byproduct of deterioration of the Fordist model of industrial mass production, which collapsed through the transformation from a manufacturing-based economy to a post-industrial service sector economy. Many contemporary urban problems are also the consequences of globalization. As Margit Mayer (1991) points out, the Fordist growth model relied on a regulatory downtown market and capital/labor relationship. Under this model, local governments functioned as subordinate agencies while states and/or regions served as administrative units that channeled growth and distributed it evenly throughout the nation. The principal focus of municipal politics in Fordist cities was on the expansion of the urban infrastructure and the management of large scale urban renewal. But as Rae (2003) makes clear, there was little need for public intervention because the economy was self-propelled.

The economy has since polarized into two different but international growth sectors. At one extreme is the high-paying corporate service sector and at the other is the low-paying sector of downgraded manufacturing and lower level services. This only defies the class composition of urban populations. The crisis in Fordist accumulation wiped out huge numbers of formerly stable manufacturing jobs. While urban space is now shared functionally and economically, it is also socially segregated and culturally

differentiated. Another effect of the post-Fordist restructuring is competition. With capital no longer geographically rooted, localities now compete with one another, with the consequence that so-called world cities have become players in the world economy. Whereas the primary business of Fordist cities was to implement, administer and filter decisions that were made above at the national and regional level, the new post-Fordist cities have been forced to develop more entrepreneurial strategies (Mayer 1991).

During the 1950s and 1960s, international trade, capital investment, and labor migration patterns contributed to economic growth in the Western industrial nations. These countries in turn would export to developing nations and would also import low-cost energy, as well as raw materials. All this began to change by the 1970s (Judd and Parkinson 1990). During the post-World War II period, industrialization appeared to offer the key to economic prosperity. Cities with large and diverse manufacturing bases promised secure growth and stable employment. Globalization in recent years has only enhanced the importance of financing, informational and control functions, but it has also enlarged the number of competitors in the tertiary sector. Consequently, localities are always having to adjust to forces beyond their control. For many medium size cities, the weakening in importance of their natural advantages has meant termination of their *raison d'être*.

Historically, urban processes were defined by what James O'Connor (1973) referred to as an implicit partnership between public and private sectors, in which it was presumed that capitalists would make prudent investments that would generate jobs and economic opportunities, while the public sector would make capital and social investments—education, job training, infrastructure—so as to ease the burden of capitalists. With the fall

of the Fordist model, public officials, especially at the local level have become what Peter Eisinger (1988) essentially calls entrepreneurs. No longer is local government a passive cheerleader who creates an environment conducive to investment—the “favorable business climate—but an active participant who courts it. Local public officials have essentially become salespeople who sell their communities as good places to do business.

Urban Theory

Urban theory might best be defined by Kevin Cox's (1983) suggestion that local governments attempt to attract utility enhancing activity, which is usually defined in fiscal terms. In the United States, a major policy objective of local government is the minimization of tax rates. But to achieve this objective, it is necessary to maximize the tax base relative to expenditures, which means they need to both attract individuals and associated land uses that will yield positive fiscal externalities and exclude those that result in negative fiscal externalities. In other words, attracting that which enhances the tax base while excluding that which weakens it in essence defines urban political activity, and is at the core of urban political economy because urban political processes are very much dependent on urban, national, and now global economic processes. Urban development and continued sustainability require growth, and this reality drives urban policy in most cities around the world.

Harvey Molotch (1976) long ago characterized cities as growth machines, whereby the principal function is to generate economic growth and to in turn pursue policies that will facilitate growth. Politically, this results in public officials forming coalitions with business leaders for the purposes of generating growth oriented policies. Not all constituents, however, will

necessarily be on board. Therefore, local politicians in an effort to appear representative of all interests, and not just business interests focus on a host of “symbolic” issues such as crime, education, and welfare for the purpose of mass consumption. Symbolic issues, however, divert public attention from the fact that local officials may be in coalitions with so-called outside interests for the sake of the supposedly larger community interest. The urban regime, then, is one of economic growth, and it is this need for growth that ultimately defines the character and composition of the regime. Nevertheless, urban growth coalitions increasingly find themselves in a prisoner’s dilemma, whereby the success in funding new areas of specialization depends on leadership groups elsewhere not initiating the same strategies (Fainstein 1990).

In a variant of the growth machine theory, Paul Peterson (1981) argued that urban policy is in large measure determined by the need to attract investment and that urban politics are effectively constrained by the need to make cities appealing to would-be investors. Peterson identifies three broad policy areas: development, redistribution, and allocation. Development policies essentially enhance the economic position of cities. Redistributive policies are of benefit to low-income residents, but at a cost to the local economy. And allocational policies are more or less neutral with regards to their economic effects. Public officials are loathe to pursue allocational policies, for fear that they could adversely affect investment decisions, thereby creating an unfavorable business climate. So long as urban economies were growing and expanding, local governments were in a position to spend on programs that would have distributional/allocational benefits.

Cities need to generate development, and to do this they must attract investment dollars.

As Peterson explains: “The interests of cities are neither a summation of individual interests nor the pursuit of optimum size. Instead, policies and programs can be said to be in the interests of cities whenever the policies maintain or enhance the economic position, social prestige, or political power of the city, taken as a whole” (p.29).

Local officials are likely to be sensitive to the economic interests of their communities for three basic reasons. First, economic prosperity is essential to protecting the city’s fiscal base. The health of the fiscal base has to stem from economic growth; it cannot be a function of greater taxation because that could have the adverse effect of chasing people away. Second, good government is good politics. It will ultimately be to the advantage of local politicians to pursue policies that contribute to the economic prosperity of the local community as a whole. And third, local officials may have a sense of community responsibility. The economic well-being of the community affects the health of local businesses and ultimately the fate of workers.

All these things together affect land values and the city’s cultural life. Therefore, in seeking to posit the primary urban interest, Peterson suggests that it would be the maintenance of their economic vitality. Peterson’s typology captures well the dilemma faced by many cities, and ultimately serves as a framework for addressing urban issues. Because most policy problems tend to be concentrated in urban areas, they by extension form the core of urban and regional problems. Such problems include unemployment, redevelopment, housing, education and training, crime, welfare and welfare reform, and the various social pathologies plaguing American cities, and others around the world.

Paul Kantor (1995) characterizes cities as “dependent” in that they have no choice but

to pursue development policies due to changing economies. This notion of dependence has only led to Regime theory, that cities are governed less by formal political structures, and more by informal regimes in which formal structures are contained. Regime theory suggests that urban politics can be characterized by the presence of coalitions who are actively involved in the policymaking process. Clarence Stone (1989) defines an urban regime as “the informal arrangements by which public bodies and private interests function together in order to be able to make and carry out governing decisions.” These governing decisions are not a matter of control, but of managing conflict and making adaptive responses to social change (Stone 2001). Stone defines urban regimes in general terms as consisting of a set of arrangements through which a community is governed. In regime theory, the political system is characterized by a division between the market and the state. Whereas the state is responsible for public and legal powers and the overall welfare of its citizens, the market is where productive assets are concentrated, usually in private hands. And yet, to govern effectively, i.e. to make public policy, requires that this division exist. In regime theory, the political elites essentially come together with the economic elites (Ferman 1996).

While cooperation is essential in this relationship, it is by no means automatic. Andrew Merrifield (2002) suggests that the city’s unique position and its consequent need for growth lends itself to what he refers to as “dynamic urbanism”, which perhaps best characterizes contemporary urban politics. A dynamic urbanism ultimately results in a dialectic between forces of development—and development out of necessity—and forces of protest. For every action taken by the regime, there may be an opposing political reaction. It then becomes the task of

formal political leaders to balance these contending forces. What these forces of protest then accomplish is to make the regime perhaps more accountable to all urban interests. But this dialectic effectively defines the political nature of most urban systems.

Effects of Globalization?

Urban form has always been determined in large measure by economic forces. The industrial city looked different from the pre-industrial city. What, then, is the effect of globalization? Marisa Carmona (2000) suggest that the basic form of urban development is increasingly being determined by three fundamental processes: globalization, environmental change, and the changing relationship between the states and the civic societies. The dependency that Kantor identified has become the cornerstone of urban centers worldwide, as they have become dependent on international finance and investment. As Sako Musterd and Wim Ostendorf (1998) note, central concepts in contemporary urban debates include segregation, social polarization, and social exclusion. A city’s economic structure, and the type of restructuring taking place are frequently viewed as among the most powerful forces behind social fragmentation and integration in the urban realm. Over the past decade, advanced industrial countries have all gone through a process of economic restructuring assumed to be strongly associated with the process of globalization. The position of immigrant groups relative to the position of indigenous populations appear to vary by states.

Chris Hammett (1998) notes that the extent and forms of social polarization in different countries is unlikely to be homogenous or unidirectional. Rather they are the consequence of economic restructuring which has resulted in changes in the structure of the labor market, the structure of occupations and

incomes and the division between those who are actively employed and those who are not. Andrew Beer and Clive Forster (2002) investigate the relationship between economic restructuring, social polarization and programs and policies of the Australian government. Australia's major cities have been profoundly affected by economic restructuring. Australia is a highly urbanized country, with more than 85 percent of the population living in urban areas in 1991. At the same time, Australian cities have generally been low density and highly suburbanized. During the 1980s and 1990s, the Australian government embraced international processes of economic change that were transforming the economy and sought to hasten the emergence of the new economic order. Income inequality increased as income of the poor fell between 1976-1981, but after 1981 it increased as the wealthy became wealthier. While welfare policies pursued by both the Hanke and Keating Labor Governments muted the impact of economic restructuring, they failed to prevent the emergence of urban regional disadvantage.

Sao Paulo, for instance, with a population of 10 million in its municipality and 17 million in its larger metro area is described as a viable city, but is also becoming more and more dependent on the comings and goings of international finance and investment-flows (Rocco 2000).

In Europe, cities have generally been experiencing a growing problem of social exclusion, aggravated by spatial segregation, especially concentrated among disadvantaged groups. These disadvantaged include the unemployed, the young and the unskilled. As much as there may be any number of explanations for social exclusion, a common underlying factor is change in economic structure, stemming from global competition and technological innovation. Paul Slouten

(2000) notes that in Dutch cities there is under-representation of unskilled workers—the most rapidly declining group—while at the same time an increase in jobs requiring greater skills. And yet, in Western societies polarization tends to be mediated by the structure of welfare provision and taxation. In American cities, polarization may owe to a specific institutional context, mainly the high and growing level of immigration, and its implications for labor supply. According to Jerome Kaufman (1998), Chicago is still considered to be one of the world's major cities in the post-industrial era. About 2/3 of the city's population are minorities, with African-Americans, the largest minority, comprising 38 percent of the city's population. The city is highly segregated, and it is one of the most segregated in the U.S. Its job losses have also been considerable: from 1963-1982, 269,000 manufacturing, 64,000 retail, and 47,000 wholesale jobs were lost, while only 57,000 selected service jobs were gained. The new urban poverty is essentially characterized by segregated neighborhoods inhabited by poor blacks, in which a substantial majority of individual adults are either unemployed or have simply dropped out of the labor force. So extreme is the segregation of blacks that it is strongly believed that it has greatly hindered opportunities for moving up the social mobility ladder. But at the same time, it has always been a city of immigrants.

Metropolitan Toronto too is also a major reception area for Canada's immigrant population, and is one of the most ethnically diverse metropolitan areas in the country. The life chances of immigrants, and also native-born Canadians, are largely determined by the various policies of Canada's welfare state. In global terms, Canada is considered to be a liberal welfare state, but at the local level there tends to be an uneven delivery of services. During the last three decades, there

has been a shift in origins of Toronto's immigrants. Those immigrants flowing into the city represent a wide spectrum of economic groups. They include refugees admitted on humanitarian grounds; those joining families already in the country; business people with entrepreneurial skills and capital to invest; and independents admitted through a point system or persons with relatives in Canada who agree to provide financial support if needed (Murdie 1998).

Sako Musterd and Wim Ostendorf (1998) note that in Amsterdam, social exclusion related to the lack of social participation is one of the most threatening problems for city officials. Social exclusion, then, is one of the most important potential consequences of many of those processes related to social problems. In many Dutch cities spatial patterns can be labeled 'mosaics' rather than 'polarized entities'. In the Dutch situation, spatial segregation of poverty has had limited or no influence on the social participation of the population. In Sweden, however, disadvantaged people tend to cluster voluntarily or involuntarily, in isolation from mainstream social and economic activities. Addressing income inequality has long been an overarching political goal. The government has also attempted to mix different groups of households in 'integrated housing'—ideally a mix of households with different demographics, socio-economic and ethnic characteristics. Swedish welfare policy has also been focused on economic resources. An ideological cornerstone of the Swedish welfare state has been equality between households, despite demographic, socio-economic and ethnic characteristics, as well as residential patterns (Borgegard et al 1998).

Regional Governance

Because of the unique position of cities, various governments have experimented with regional governance. During the late 1940s,

there were calls in the Toronto metropolitan region, for instance, to enhance economic prospects and stem the physical decline of inner-city neighborhoods. Metropolitan analysts in the U.S. were already observing that some municipalities gain from metropolitan area expansion while others lose. Metro's planning principles were premised on the need to control sprawl, protect economic and cultural vitality of the central city, provide a balance of residential and employment opportunities in all parts of the area and exchange development at densities high enough to support public transit systems.

Frances Frisken (2001) observes that the 1953 creation of a federated form of metropolitan government for the city of Toronto and its 12 surrounding municipalities was in sharp contrast with many failures to achieve similar metropolitan reform in the United States. The creation of Metropolitan Toronto, however, was only the first in a succession of changes in Toronto regional governance as the region evolved. Others included reorganization that brought suburban Metropolitan Toronto into four two-tier regional municipalities. Subsequently in 1998, the governments of Metro and its member municipalities were reorganized into an enlarged version of Toronto, the Greater Toronto Area (GTA). But it wasn't simply a question of these subunits getting together to form a larger metropolitan unit, rather the Ontario government was actively involved in the process. As she notes, the Ontario government did more than restructure governments; it initiated local revenue sharing for a variety of municipal purposes. It initiated a fully integrated transit system linking the core city and its suburbs, but only within Metro. It initiated efforts to achieve or encourage a fair share distribution of low-rental housing and equality of educational funding. And it legislated context for many

municipal and regional planning and land use regulation.

In Korea, for instance, K. Hong (1997) notes that policy has sought to achieve a more balanced geographical distribution of income, increased opportunities, and access to public services. South Korea has specifically had to address regional disparities in income between both urban and rural areas and between metropolitan and non-metropolitan regions. From the early 1960s, the Republic of Korea experienced rapid economic development and transformation from an agriculture-based economy to a modern industrial one, with the result being a shift in population and economic activity within the country. One cause of regional inequality may have been due to central government policies aimed at maximizing national development in the interests of efficiency. Moreover, the tendency toward polarization may have been exacerbated by the concentration of public investment in a few areas, including the Seoul and Pusan metropolitan regions. And as Korean society continues to evolve against the background of rapid economic growth, its regional problems are likely to become increasingly similar to those of developing countries.

The focus of regional policy has essentially been on the issue of regional imbalance—decentralization of population and economic activities in Seoul and large metropolitan areas. Major instruments of regional policy have been: 1) directing growth of Seoul with the re-zoning of industrial land use and the relocation of manufacturing industries in violation of city ordinances; 2) redistribution of industrial activities to changing regions; 3) relocation of central government functions to local areas; and 4) the stimulation of regional economies, particularly in provincial areas. As Hong suggests, the main function of the central government should be to coordinate policies between regions, and that it is crucial

to strengthen local self-government if the self-sufficiency and economic development capacity of localities is to be raised. Moreover, if the central government is to finance economic development at the local level, a necessary first step would be to redesign the mechanism for allocating intergovernment-financial resources, and to improve the channels of resource allocation to local government authorities.

Meanwhile, the transformation of Central and Eastern Europe has created new spatial patterns of economic and social inequality. John Bachtler and Ruth Downes (1999) observe that the capital cities have been flourishing. Unemployment has been relatively low and new businesses have been created, and foreign investment has been especially concentrated in these cities. Meanwhile, the older industrialized areas have suffered from the closure of outdated inefficient enterprises. The uneven spatial impact of internal economic reform is becoming apparent as the transformation proceeds, and the overall picture is one of widening disparities between and within countries. In all the transition countries, the major agglomerations and urban centers lead in the transformation process, and the most advanced features of transition are evident in urban areas, with foreign investment heavily concentrated in the cities. The attractiveness of major urban areas for foreign investment combined with privatization of housing has caused a steep rise in rents and a re-stratification in social groups within cities, thereby resulting in a new residential geography. And the challenge for institutional structures throughout Central and Eastern European countries following these political and economic reforms has been the reorientation to market economic requirements.

Similarly, as Blazyca et al (2002) observe, an enduring feature of economic development

in Poland has been geographic imbalance. Regional policy in communist Poland was targeted towards 'equalization'—its principal instrument of investment. If centers believed that it had to do something for weaker regions, it would be enough to take new plants to towns concerned. And during the period immediately following the collapse of communism in 1989, macro-economic stabilization became the priority. As Poland was being spontaneously reshaped by market forces, its regions were left to benign neglect. A regional policy, then, was intended to help secure this transformation process by helping to develop a market economy in all regions also ensuring that all regions will provide employment opportunities, basic services and a clean environment. A regional policy would also be designed to reduce regional disparities, especially between core and periphery, east and west, and urban and rural.

Leadership and Rejuvenation

The scale of economic decline in both European and North American cities has provoked debate since the 1970s about the future of cities, and whether they can be and should be "saved (Judd & Parkinson 1990)." A significant number of these cities have arrested this decline and even experienced some economic recovery. What, then, are some of the approaches to addressing various urban problems? According to Rae (2003) this decline has indeed signaled the end of urbanism, the process by which cities grew correspondingly with industrialization. Using New Haven, Connecticut as an example, during the period of growth, City Hall was marginal to the city's economic life. Local government provided virtually no housing and built no malls or office towers. All it did was maintain an infrastructure, maintained public safety, and exercised police powers. Rather urban society was relatively self-regulating, and local government didn't need

to intervene too often. In New Haven, the economic governance of the city was effectively delegated to the private sector. With the end of urbanism, however, local officials have had to do more to maintain the economic viability of their cities.

In the U.S., as well as other countries, local officials have actively entered into partnerships. One of the most famous partnerships was the Poletown case involving a partnership between the City of Detroit and General Motors (GM). Detroit had long been in a state of economic decline and GM was now willing to build two new Cadillac plants in Detroit under certain conditions. GM needed a suitable site and had asked Detroit's Mayor, Coleman Young, to help locate a site that would be suitable to GM's needs. After searching, the site located was a neighborhood known as Poletown. GM would construct these plants and in the process create 6000 new jobs. GM was even prepared to pay the city \$18,000 an acre, but the city had to acquire the property, which would need to be done through the state's power of eminent domain and then prepare the site, which consisted of demolishing existing structures. The cost to the city, then, was to be \$200 million, for which the city would need both state and federal assistance. Additionally, the city would have to provide GM with tax abatements. Although GM was prepared to invest \$600 million in the City of Detroit, local officials would also need to make a public investment of \$250 million (Jones & Bachelor 1993:80-83).

Nevertheless, Poletown was controversial because the construction of these two plants and their supposed benefits to the city were to require no less than the wholesale destruction of a community. This meant the demolition of homes, schools, churches, cemeteries and other social and civic institutions. The Poletown community was essentially a poor community that was also divided along racial,

ethnic, religious, and life-cycle lines. The black community was divided between young and middle-aged families with children. Members of the Polish community were uniformly older and nearing retirement age. Most of this latter community had actually lived in the neighborhood for more than 20 years and had developed a strong sense of neighborhood belonging. And what opposition there was to this project came mainly from the long-term Polish homeowners who had a strong sense of neighborhood and community. Still, many residents were not necessarily opposed to the project. The city had actually moved aggressively to preempt neighborhood opposition by offering each resident generous relocation payments. Moreover, not only was the project supported by local officials, but it had the backing of labor, church, and other civic leaders. This was regime theory in practice. In the end, however, G.M. closed those plants down several years later, which only raised the question of to what lengths should local officials go to generate growth. And yet, it would also appear that local officials were able to take advantage of political segregation created by the economic decline to create sufficient dissension among the ranks of their populations in order to accomplish their policy objectives.

As much as cities are subject to outside forces beyond their immediate control, their ability to adapt and maintain themselves as viable urban centers is very much a function of leadership. And much leadership has revolved around the development of high-tech industries as substitutes for the older deteriorating manufacturing bases. In Hamburg, Germany, for instance, the first attempt to improve the city's economy following World War II was during the mid 1960s, and was characterized by the city elite's general failure to understand changing international economic processes. The most

important principle of economic policy during this period was to increase the city's industrial potential by offering new industrial spaces for existing and new firms. More recently, however, industrialization policy has revolved around the creation of a milieu that would stimulate technological innovation. The roots of this policy lay with the decision to establish a Technical University in Hamburg in the late 1970s. Located in the southern district of the city, it was intended to achieve two goals: to close the economic gap between the northern and southern parts of the city; and to develop the economy of the southern region in general, and to support small and medium sized firms by cooperating in research and production. The city's renewal and improvement programs of inner-city areas prior to 1983 were designed to slow down suburbanization that had led to population losses of more than 15,000/year during the early 1970s. And yet, the social and economic problems faced by Hamburg mirror those faced by comparable old industrialized cities (Dangschat & Ossenbruegge 1990).

Paul Lawless (1990) suggests that in the English city of Sheffield what was central to the political debate during the early 1980s was economic decline and employment contraction. The local economy also proved to be a major focus of policy development. In 1981 the city created the Employment Department, but more importantly, it strove to create partnerships, which were driven by the Chamber of Commerce. This informal relationship between the city and the Chamber of Commerce eventually led to the creation of the Sheffield Economic Regeneration Committee (SERC) in 1987. The Committee was to consist of representatives from a range of local interests, including the authority, the private sector, trade unions, community groups, higher education, and the Sheffield Development

Corporation. The emergence of the SERC falls into the broader rubric of growth coalitions, and the city is illustrative about key issues concerning growth coalitions in Britain. That is, the local authority has been a key driving force behind growth initiatives, the range of interests involved in Partnerships is not great, and property interests are not especially significant.

Similarly in Montreal, business organizations have been involved in major public-private efforts during the past decade. Since the end of the 1970s, Montreal's policies on revitalization consisted of four components: manufacturing industries, neighborhood commerce, housing construction, and tertiary functions in the central business district (Leveillee & Whalen 1990). The point, as Judd et al (1990) make clear is that leadership is the crucial variable in determining how cities respond to economic change. Leadership capacity, however, is affected by the skills with which leaders are able to exploit available resources. Those cities that have been able to develop substantial leadership capacity have also been able to develop more complex strategies for regeneration than those cities whose leadership is weak. And yet, what is clear is that the leadership being exhibited here is active leadership. As cases like Poletown demonstrate, local officials who are able to seize upon opportunities as they arise may be able to achieve some success, if albeit temporary success. And yet, a willingness to enter into such partnerships—to adjust to new realities—also sends signals to would be investors that such localities are indeed places to do business.

Conclusion

The evolution of cities is such that we need to think of city governance more broadly than simply local officials forced to attract investment and generate growth. These issues

that are traditionally regarded as urban are no longer merely local, but are more broadly regional, and increasingly becoming national. To a certain extent, globalization highlights the reality that cities, and their form, are very much contingent on outside forces, and continuously they need to be adaptable. Partnerships between public and private actors today appear to be very similar to what they were in the past. Not only do political leaders have to create environments conducive to favorable investment by not pursuing policies that would create unfavorable business climates, they have to actively court it. The concept of regimes becomes even more important because it best characterizes the typical urban environment. On the one hand, the current political economy of urban areas highlights that cities, and more broadly their surrounding metropolitan areas, are at best autonomous. It underscores that their political activity is in fact driven by outside forces. On another level, it may call attention to the fact that local units—cities per se—may no longer be viable units of analysis.

Selected References

- Alonso, William. (1983) "A Theory of the Urban Labor Market." in Robert W. Lake ed. *Readings in Urban Analysis: Perspectives on Urban Form and Structure*. New Brunswick, NJ: Center for Urban Policy Research/Rutgers University.
- Bachtler, John and Ruth Downes. (1999) "Regional Policy in the Transition Countries: A Comparative Assessment." *European Planning Studies*, Volume 7, Number 6, pp. 793-808.
- Beer, Andrew and Clive Forster. (2002) "Global Restructuring, the Welfare State and Urban Programmes: Federal Policies and Inequality within Australian Cities." *European Planning Studies*, Volume 10, Number 1, pp. 7-25.

- Blazycza, George, Krystian Heffner, and Ewa Helinska-Hughes. (2002) "Poland—Can Regional Policy Meet the Challenge of Regional Problems?" *European Urban and Regional Studies*, Volume 9, Number 3, pp. 263-276.
- Borgegard, Lars-Erik, Eva Andersson and Susanne Hjort. (1998) "The Divided City?: Socio-Economic Change in Stockholm Metropolitan Area, 1970-94." in Sako Musterd and Wim Ostendorf (Editors), *Urban Segregation and the Welfare State: Inequality and Exclusion in Western Cities*. London & New York: Routledge.
- Carmona, Marisa. (2000) "IBIS Research Framework Development." in M. Carmona and J. Rosemann (Editors), *Globalization, Urban Form and Governance: Second Industrial Conference ALFA-IBIS Proceedings*. The Netherlands: Delft University Press.
- Cox, Kevin R. (1983) "Local Interests and Urban Political Processes in Market Societies." in Robert W. Lake (Editors), *Readings in Urban Analysis: Perspectives on Urban Form and Structure*. New Brunswick, NJ: Center for Urban Policy Research/Rutgers University.
- Dangschat, Jens S. and Juergen Ossenbruegge. (1990) "Hamburg: Crisis Management, Urban Regeneration, and Social Democrats." in Dennis R. Judd and Michael Parkinson (Editors), *Leadership and Urban Regeneration: Cities in North America and Europe*. Newbury Park, CA: Sage Publications.
- Eisinger, Peter K. (1988) *The Rise of the Entrepreneurial State: State and Local Development Policy in the United States*. Madison: University of Wisconsin Press.
- Fainstein, Susan. (1990) "The Changing World Economy and Urban Restructuring." in Dennis Judd and Michael Parkinson ed. *Leadership and Urban Regeneration: Cities in North America and Europe*. Newbury Park, CA: Sage Publications.
- Ferman, Barbara. (1996) *Challenging the Growth Machine: Neighborhood Politics in Chicago and Pittsburgh*. Lawrence: University Press of Kansas.
- Friskien, Frances. (2001) "The Toronto Story: Sober Reflections on Fifty Years of Experiments with Regional Governance." *Journal of Urban Affairs*, Volume 23, Number 5, pp. 513-541.
- Hammett, Chris. (1998) "Social Polarization, Economic Restructuring and Welfare State Regimes." in Sako Musterd and Wim Ostendorf ed. *Urban Segregation and the Welfare State: Inequality and Exclusion in Western Cities*. London & New York: Routledge.
- Hong, K. (1997) "Regional Policy in the Republic of Korea." *Regional Studies*, Volume 31, Number 4, pp. 417-423.
- Jones, Bryan D. and Lynn W. Bachelor. (1993) *The Sustaining Hand: Community Leadership and Corporate Power*. Lawrence: University Press of Kansas.
- Judd, Dennis and Michael Parkinson. (1990) "Urban Leadership and Regeneration." in Dennis Judd and Michael Parkinson ed. *Leadership and Urban Regeneration: Cities in North America and Europe*. Newbury Park, CA: Sage Publications.
- Kantor, Paul. (1995) *The Dependent City Revisited: The Political Economy of Urban Development and Social Policy*. Boulder, CO: Westview Press.
- Kaufman, Jerome. (1998) "Chicago: Segregation and the New Urban Poverty." in Sako Musterd and Wim Ostendorf ed. *Urban Segregation and the Welfare State: Inequality and Exclusion in Western Cities*. London & New York: Routledge.
- Lawless, Paul. (1990) "Regeneration in Sheffield: From Radical Intervention to Partnership." in Dennis R. Judd and Michael Parkinson ed. *Leadership and*

- Urban Regeneration: Cities in North America and Europe*. Newbury Park, CA: Sage Publications.
- Leveille, Jacques and Robert K. Whalen. (1990) "Montreal: The Struggle to Become a World City." in Dennis R. Judd and Michael Parkinson ed. *Leadership and Urban Regeneration: Cities in North America and Europe*. Newbury Park, CA: Sage Publications.
- London, Bruce and William G. Flanagan. (1976) "Comparative Urban Ecology: A Summary of the Field." in John Walton and Louis H. Masotti, ed. *The City in Comparative Perspective: Cross-National Research and New Directions in Theory*. New York: Sage Publications/John Wiley & Sons.
- Mayer, Margit. (1991) "Politics in the Post-Fordist City." *Socialist Review*, Volume 21, Number 1, pp. 105-124.
- Merrifield, Andrew. (2002) *Dialectical Urbanism: Social Struggles in the Capitalist City*. New York: Monthly Review Press.
- Molotch, Harvey. (1976) "The City as Growth Machine." *American Journal of Sociology*, Volume 82, Number 2, pp. 309-355.
- Murdie, Robert A. (1998) "The Welfare State, Restructuring and Immigrant Flows: Impacts on Socio-Spatial Segregation in Greater Toronto." in Wim Ostendorf ed. *Urban Segregation and the Welfare State: Inequality and Exclusion in Western Cities*. London & New York: Routledge.
- Musterd, Sako and Wim Ostendorf. (1998) "Segregation, Polarisation and Social Exclusion in Metropolitan Areas." in Sako Musterd and Wim Ostendorf ed. *Urban Segregation and the Welfare State: Inequality and Exclusion in Western Cities*. London & New York: Routledge.
- Musterd, Sako and Wim Ostendorf. (1998) "Segregation and Social Participation in a Welfare State: The Case of Amsterdam." in Sako Musterd and Wim Ostendorf ed. *Urban Segregation and the Welfare State: Inequality and Exclusion in Western Cities*. London & New York: Routledge.
- O'Connor, James. (1973) *The Fiscal Crisis of the State*. New York: St. Martin's Press.
- Peterson, Paul E. (1981) *City Limits*. Chicago: University of Chicago Press.
- Rae, Douglas W. (2003) *City: Urbanism and its Ends*. New Haven: Yale University Press.
- Rocco, Roberto. (2000) "Sao Paulo: Globalization, Governance and Urban Form." in M. Carmona and J. Rosemann ed. *Globalization, Urban Form and Governance: Second Industrial Conference ALFA-IBIS Proceedings*. The Netherlands: Delft University Press.
- Stone, Clarence N. (2001) "The Atlanta Experience Re-Examined: The Link between Agenda and Regime Change." *International Journal of Urban and Regional Research*, Volume 25, Number 1, pp. 20-34.
- Stone, Clarence N. 1989. *Regime Politics: Governing Atlanta, 1946-1988*. Lawrence: University Press of Kansas.
- Slouten, Paul. (2000) "Strategic Planning, Sustainability, and Urban Regeneration." in M. Carmona and J. Rosemann ed. *Globalization, Urban Form and Governance: Second Industrial Conference ALFA-IBIS Proceedings*. The Netherlands: Delft University Press.
- Walton, John and Louis H. Masotti. (1976) "Comparative Urban Research: The Logic of Comparisons and the Nature of Urbanism." in John Walton and Louis H. Masotti, ed. *The City in Comparative Perspective: Cross-National Research and New Directions in Theory*. New York: Sage Publications/John Wiley & Sons.
- Warner, Sam Bass. (1987) *The Private City: Philadelphia in Three Periods of its*

Growth. Philadelphia: University of
Pennsylvania Press.

Oren M. Levin-Waldman
School of Public Affairs & Administration
Metropolitan College of New York
New York City, USA
OLevin-Waldman@mcny.edu

Welfare State

Anne de Bruin

Introduction

In capitalist economies, welfare may be sourced from the state (public), market, and non-market (includes the family and the voluntary third sector) sectors. The share of each of these sectors in welfare provision determines the welfare mix. This welfare mix varies in different time periods and across countries. In the traditional, non-monetised society, the family had a monopoly on welfare provision. The spread of monetisation, particularly with the Industrial Revolution, led to the emergence and consolidation of the mixed economy of welfare. The welfare state was “the latest stage in a dynamic process of adjustment between individual and society (Fraser 1973:222). Stanfield and McClintock highlight the welfare state as “something of a social and economic hybrid”, representing the blend of public and private involvement in the economic process (1999:1245).

There is no unambiguous definition of the welfare state. Writing at a time which might be described as the heyday of the welfare state, Titmuss referred to it as an “indefinable abstraction” with which he was “no more enamored today” than he had been two decades earlier (1968:124). Barr points out that “the welfare state is one of those concepts that defies precise definition ... the boundaries of the welfare state are not well defined” (1993:2-3).

Often, the term welfare state is used generically to refer to public programmes, chiefly in the areas of income maintenance, health, education, training, social welfare services and housing. These programmes are, however, only the more obvious manifestation of the principle of the modern Western welfare state, which is, that the state

has the ultimate responsibility to ensure that the material well-being of any of its citizens does not fall too far below that of the average citizen. There can, nevertheless, be varied normative interpretations of what this minimum level of well-being should be. For example in New Zealand, the 1972 Royal Commission on Social Security (RCSS) formalised this minimum level in terms of the state ensuring a standard of living that would allow “*participation in and belonging to the community*” (RCSS 1972:65, original italics).

In attempting to construct a “general theory of the welfare state”, Spicker (2000) deliberates basic propositions which are developed through a series of sub-propositions. “The welfare state is a means of promoting and maintaining welfare in society” constitutes a core proposition with follow-on sub propositions which include, “The welfare states provide social protection” and “welfare is promoted and maintained through social policy.” Despite a broad perspective on what constitutes the welfare state as conveyed through the basic proposition, the sub-proposition exposition, clarifies a narrower, yet commonly adhered to interpretation of the welfare state – one which has a social policy and social protection systems focus. The welfare state discourse can thus be conducted both in terms of this narrower perspective, as well as at a broader underpinning level of the role of the state.

Origins and Historical Context

Generally there is consensus that the modern welfare state was initiated when governments enacted social security legislation to establish government funded income security measures. These represented a clear break with the earlier poor relief forms which afforded a degree of social protection but at the expense of citizenship rights. Thus the introduction of governmental programmes for worker insurance in Bismarck’s Germany in

the 1880s, was among such first steps toward state provided welfare.

The term 'welfare state' is thought to have been coined by William Temple, the Archbishop of York, when he distinguished between the "welfare state" serving the common interests of citizens and the "power state" which serves the interests of tyrants as in Nazi Germany (Temple 1941; cited in Barr & Whynes 1993:6). Losing its original religious and moralistic overtones, the term became popularised in Britain after the Second World War and its use spread to other developed capitalist economies. It came to be used as a convenient way of referring to the economic and social policy changes that were taking place at the time. These policy changes had three broad strands: the introduction and extension of state provision of social security, health, education, housing, employment and other welfare services; the maintenance of full employment; and a programme of nationalisation. Together these strands constituted the welfare state (Johnson 1987:3).

At an overarching level, the historical construction and basis of the welfare state may be viewed in terms of different phases of capitalist development. The restructuring of welfare states in the last three decades can be set against the background of a movement of capitalist economies to a new phase of development, a new socio-economic and global era. A variety of labels, such as, fifth Kondratiev cycle or long wave, post-Fordism, post-industrial, post-modernist, are used to describe this new phase. The new era involves changes in both the organisation of enterprises and the organisation of work, as well as the international division of labour. An understanding of the post-Fordist debate and political economy strands of this discourse is necessary in order to explain the roots of welfare state changes at national levels.

The sustained prosperity of the developed capitalist countries during the post-war period to around the mid 1970s, has been described as the 'golden age of capitalism' and the term Fordism, derived from the techniques of production pioneered in the American auto assembly factories of Henry Ford in the early 1900s, is used to describe the dominant hegemonic post-war development model of advanced capitalist economies of this age. Since Keynesian economics was the economic paradigm of the time, the progression of the modern welfare state may be viewed as a corollary of Keynesian economics. Keynesian thought underpinned the social alliance and national consensus on economic and social goals and policies that emerged in the period and the welfare state form that characterized this time may therefore be differentiated as the 'Keynesian Welfare State' (KWS).

The Keynesian Welfare State

An important aspect of the KWS was the unity of the capital-labour relation or relations between employers and workers, which prevailed. Though this capital-labour relation model had many national brands, it typically comprised four major strands: large-scale division of labour; specialisation and mechanisation of manufacturing; mass production of standardised goods, and rather strong unions (Boyer 1995:23-24). It gave rise to increases in productivity and a compromise on productivity sharing, which influenced wage formation. Conceptually, after the Second World War the wage was no longer a pure market variable since it took into account a minimum standard of living. This wage was then raised according to the general advances in productivity (Boyer 1995:25). In New Zealand for instance, the wage level, buttressed by full employment, guaranteed the male breadwinner the ability to comfortably maintain himself, a wife and

three children. The welfare state was a “wage-earner’s welfare state” (Castles 1985) and working class pragmatism sought to procure their welfare through state support of wage and employment security rather than through more comprehensive forms of welfare as in the Scandinavian welfare states.

The KWS was characterised by policy commitment to the full employment, seemingly made possible by the pursuit of Keynesian economic management. The San Francisco United Nations Charter, drawn up at the close of the Second World War, had included the clause under Article 55 that “the United Nations shall promote: higher standards of living, full employment and conditions of social progress and development”. Even in West Germany, where much emphasis was placed on price stability, the goal of full employment was explicitly declared in the Stabilisation Law of 1967, at quasi-constitutional level. The KWS was also based upon the stable nuclear and patriarchal family form, with full employment therefore primarily male full employment.

Neoclassical, monetarist and public choice economic theory provided an analytical rationale and the stagflation of the post oil shock world economy supported a new direction in economic and social policy and the demise of Keynesian stabilisation and regulation of the market economy. By the mid 1980s, the notion that the KWS was in “crisis” was widely accepted and received prominence in the literature (see, e.g., OECD 1981). According to Galbraith, the discarding of the KWS was cemented by “A Contented Electoral Majority” or a “Culture of Contentment” and its line of economic thought (Galbraith 1992). Serving this contentment are three basic requirements: limited government intervention, social justification for the uninhibited pursuit of wealth, and a reduced sense of public responsibility for the poor (Galbraith

1992:96-97). Changing demographics and family forms - with the nuclear family no longer necessarily the norm, population ageing in affluent societies and changing gender roles in the labour market, also contributed to the erosion of the foundations of the KWS.

Implications of Globalisation

The continuation of the traditional social democratic strategy, namely Keynesian style policy, became less feasible, if not impossible, as new forces of globalisation and increased international competition took hold after the end of the post-war long cycle of world economic growth which followed the oil crisis of 1973. The “logic of globalisation” has meant the collapse of full employment, growing unemployment and economic inequalities, downward pressures on social protection and expenditure and “conflict with the ‘logic’ of national communities and democratic politics” (Mishra 1999:15).

Growing internationalisation of the labour market, coupled with the perfect (or near perfect) internationalisation of the capital market as well as the liberalisation of product markets, has altered dramatically the capital-labour model that was a feature of the KWS. A push to enhance labour market flexibility which is now frequently seen as a prerequisite for meeting the challenge of international competitiveness, has resulted in accelerated deregulation of labour markets, less centralised wage bargaining and government intervention in the bargaining process and lower levels of unionisation. High and persistent unemployment levels with labour market disadvantage of ethnic minorities and regions and the diversification of forms or employment including the growth of non-standard (atypical) jobs, pose additional threats to welfare (Sarfati 2002).

The outlook for the welfare state in the face of globalisation is however, not

completely bleak. An interesting recent study, whilst adopting a relatively narrow definition of the welfare state, with 'welfare state' taken to mean government spending on health, education, social security and welfare; examined the broad hypotheses in the literature on the implications of globalisation for the welfare state to conclude that the response to globalisation and "the fate of the welfare state appears to depend on institutional structures and policy decisions, rather than on an inevitable capitulation to global forces" (Bowles and Wagman 1998:336).

The global era has brought with it new risks for nations and individuals. Yet as Beck aptly puts it, "... risk definitions do not deprive us, but rather make political decisions *possible* ... political action gains influence in parallel to the *detection* and *perception* of risk potential' (1992:227). Dealing with risk however, is not new and lies at the core of welfare systems. Taylor-Gooby et al. (1999) point out, the British welfare state developed as a state focused response to the problem of the dealing with risks met in a typical life-course. The value placed on security, or in other words, protection from risk, also lay at the heart of the New Zealand welfare state. It is now necessary that the state is strategic, in the sense of recognising and perceiving the risk definitions of households and business. National governments have a vital role in putting in place the institutions and governance structures to enable their constituents cope in a changing risk environment of heightened global competition.

Welfare State and Economic Performance

In the new era of heightened globalisation, a broad critique of the welfare state is that it hinders economic performance and is a barrier to international competitiveness. There are two broad strands to this argument.

Firstly, it is asserted that countries with lower welfare costs are better able to compete in terms of productivity and product quality than those with higher welfare costs - there is a cost advantage of a low welfare burden. Secondly, welfare statism can be a productive liability.

Investigations on the relationship between the share of the national product claimed by the government (a proxy for the degree of welfare statism) and the growth of national product; and other studies on the influence of welfare statism on 'competitiveness' (for which there is no single definition), together with their own study (Pfaller and Gough 1991); shows contradictory evidence and no straightforward link between welfare statism and comparative economic performance (Pfaller et al 1991). There was no relation between welfare statism and productivity in the 1970s, though there was a significant positive correlation between the growth of social expenditure and manufacturing labour productivity. It is suggested that the latter finding can support two hypotheses. The first is that enterprises respond to rising welfare state costs by increasing efforts to enhance productivity. Second, although previously those countries that experienced a satisfactory growth of productivity had a greater inclination to extend their welfare states, in the different ideological climate of the 1980s, this willingness had disappeared. Although the authors maintain, that their analysis strengthens their confidence in the inherent competitiveness of the welfare state, the basic finding was that in the 1980s, there was a suggestive negative association between welfare statism and some indicators of economic performance. What is interesting is that the study also showed that fairly smooth co-operation between business and labour had been an important contributor to high-productivity, high-quality production. Inappropriate industrial relations were a

major competitive handicap (Pfaller et al. 1991:296). Similarly interesting was the finding that human capital formation was an important advantage for West Germany and Sweden compared to the other investigated countries - USA, UK and France.

The Swedish Economics Commission, arguing that productivity growth in Sweden has been significantly lower than the OECD average, accuses the Swedish model of resulting in institutions and structures that are an impediment to economic efficiency and economic growth because of their lack of flexibility and one-sided concern for income safety and distribution, and limited concern for economic incentives (Lindbeck et al. 1994). Indeed much has been written on economic disincentives resulting from public provision of various welfare measures and their continued viability in the light of adverse demographic change and fiscal constraint. While not ruling out the need for reform, caution must nevertheless be exercised when laying blame on the welfare state. As Atkinson (1995a) shows on the basis of aggregate empirical evidence provided by nine different studies of growth rate and social transfers, there is often mixed evidence. The interpretation of results depends on the theoretical framework used. The significant contribution of his paper is to demonstrate, through the use of specified models, that the institutional structure and features of benefits can change their impact on economic behaviour. Thus an identical amount of total spending on social transfers may have different outcomes for long run growth rate or the level of GDP, depending on the structure and conditions of entitlements. What may be issues of detail must not be ignored. This is a pertinent observation for other policy as well. For example in active labour market programmes issues of detail, namely design features,

strongly influences the effectiveness of such programmes.

Another study by Atkinson (1995b), highlights that moving to targeted benefits or private provision may replace one set of disincentives by another. The case of replacement of a state funded pension scheme with private provision, using an endogenous growth model, illustrates that while this may lead to a rise in the savings rate there is a fall in the desired growth rate of firms. When examining the proposition of whether the welfare state does necessarily impede economic growth, it might therefore be concluded as Atkinson does, that the “jury should stay out” as “there is a plausible case that can be entered for the defence” of the welfare state (1995b:730).

Welfare State Regimes

Distinguishing social policy configurations have been used to classify countries and their welfare state systems. A well-known classification of the affluent capitalist nations is that of Esping-Andersen's three “worlds” of welfare capitalism (1990), which has sparked a significant body of research and debate on the subject. Esping-Andersen (1990) groups welfare states into “socialist” – later referred to as social democratic (Norway, Sweden, Denmark, Finland, Netherlands); “liberal” (USA, Canada, Switzerland, Australia, Japan); “conservative” (Italy, France, Austria, Germany, Belgium), according to welfare state programmes.

Descriptors of the regime-types have been progressively modified, and the dimensions examined extended particularly to include labour market regulation and family policies (Esping-Andersen 1999). Thus citizenship-based universal benefits, comprehensive social insurance programmes and corporatist industrial relations and active family policy strongly supporting gender equality, mark the

social democratic world comprising mainly the Scandinavian welfare states. The liberal world is characterised by targeted, needs-based entitlements and is a residual model with poor family services, low levels of employment protection and un-coordinated industrial relations. The social insurance states, comprising Austria, Belgium, France, Germany, Italy and Japan, are those distinguished by occupation differentiated, employment-related social insurance, passive family policies based on the conventional male breadwinner model and coordinated industrial relations. Most recently, Hicks and Kenworthy (2003) have extended the Esping-Andersen analysis, chiefly to incorporate active labour market policy (training, job placement, etc.) and government employment, to identify only two varieties of welfare capitalism. Their label of “progressive liberalism” is a rearrangement of the social democratic and liberal regime dimensions into two poles of a single dimension. “Traditional conservatism” is the second dimension which is similar though a broader version of the conservative typology.

There is growing body of literature on the East Asian welfare state experience. Aspalter (2001) offers “conservative” to describe the welfare statism of this group of countries. Renaming the conservative Continental European welfare states centring around Germany, “Christian democratic and social democratic” after the political parties which shaped welfare policy, Aspalter delineates the features of conservative welfare state systems. They include a stronger emphasis on the regulative rather than the redistributive capacity of the state, relatively low welfare expenditure, a heavy reliance on market and third sector welfare provision and a preference for occupationally divided social insurance systems.

Hemerijck (2002:178) provides useful headings to examine the variability of welfare

states along several dimensions: *Eligibility and risk coverage* where access to social protection may be citizenship or needs based, work-related contribution or private contract based; *Benefit structure and generosity* with structure linked to country-specific social protection objectives of income maintenance, poverty mitigation and equity, and benefits being means-tested or universal, contribution-related etc. and varying from minimal to more generous; *Methods of financing* involves funding from general taxation, user charges, payroll contributions or a mix of these; *Service intensity* relates to social service provision through direct professional public services, the market or informal (extended) family means; *Family policy* varies from passive cash transfers premised on the traditional single breadwinner family form, to very active gender equality policies with generous child care and paid parental leave provisions; *Employment regulation* includes diverse ‘industrial rights’ considerations ranging from minimum wage laws to active labour market policies; *Logic of governance*; and *Industrial relations*. The logic of governance heading encompasses the management and delivery of welfare and employment policy and is also strongly linked to industrial relations. The degree of coordination and co-operation in national industrial relations can range from fragmented un-coordinated systems to sectoral wage bargaining and centralised coordination.

The logic and processes of governance is a pertinent factor, though infrequently given dedicated focus in comparative expositions of welfare states. Governance is increasingly important given the current emphasis on social dialogue and the partnership approach as a means of moving toward full employment. For instance Auer (2000) credits a successful system of “corporatist governance” and social dialogue as

significant to the labour market success of Austria, Denmark, Ireland and the Netherlands.

Towards New Terminology

It is argued that the term welfare state is not an appropriate descriptor and does not convey the current realities of the function of the state in the global age (de Bruin 2003). There is a need for new terminology to be developed to better convey the nature of the state and conceptualise the reconfiguration of the role of the state in this new era. The 'welfare state' descriptor is outmoded. Moreover, today the term 'welfare' has a popular connotation of something quite different from its initial meaning. Welfare, once synonymous with well-being, is now often perceived as ill-being. Welfare dependency and stigma frequently attached to the receipt of some welfare benefits, negatively clouds perceptions of the welfare state.

There have been a number of attempts to provide new terminology to reflect the shifting nature of state action. For instance, Jessop (1994) argues that a "Schumpeterian workfare state" is more appropriate to describe the state form and function of the emerging post-Fordist era. Distinctive objectives of this state form are promotion of innovation in open economies to strengthen structural competitiveness of the national economy through supply side intervention, and subordination of social policy to labour market flexibility and/or the constraints of global competition. These economic and social objectives represent a firm break with the KWS tradition, since now, international competitiveness of the economy takes precedence over domestic full employment and productivist social policy is given a higher priority than redistributive welfare rights.

The Schumpeterian workfare state can take different forms according to the strategies

adopted - neo-liberal, neo-corporatist and neo-statist forms, with a mixture of these also possible. In contrast to the neo-liberal regime where, social partnership arrangements are rejected, neo-corporatist strategies arise out of advance planning and concertation of economic decisions and activities by economic agents in order to further their own economic ends. Unlike under the KWS, where corporatist strategies aimed at the maintenance of full employment or stemmed from concerns about stagflation, neo-corporatist arrangements of the Schumpeterian workfare state are linked to the desire to promote innovation and structural competitiveness. There is also a movement away from macro level corporatist arrangements, as, for example, between the broad organisations of capital and labour, toward more selective, micro arrangements between, for instance, functionally distinct policy communities, such as health and education. Neo-statist strategies involve active intervention by the state to promote the structural competitiveness of the economy. Thus, the state acts to ensure dynamic efficiency of the industrial core, particularly by overseeing the restructuring of declining industries and through microeconomic targeting of policies toward particular sectors, chiefly in the high technology arena. Reskilling of the labour force is given high priority, as is the stimulation of innovation.

While the Schumpeterian workfare state and its variants have merit, in that it conveys the changed focus of the state, it is hard to see it as substitute terminology with similar comparable appeal to the popularised welfare state concept. The concept of the "strategic state" is offered by de Bruin (2003) as a better alternative and one which also fits with the entrepreneurship perspective of the global age, which as Audretsch and Thurik (1999) observe, has changed from the "managerial

economy” of the previous industrial era to a knowledge-based “entrepreneurial economy”.

The strategic state concept encompasses both the pursuit of chosen social goals and other newer systemic functions of the state. The rationale for government intervention, based on the systemic failure argument, is broader than the standard market failure argument for intervention. It involves a variety of policy responses to reduce systemic imperfections, such as eliminating informational failures by supplying strategic information, and removing institutional mismatches and organisational failures within systems of innovation. The strategic state is a key driver of innovation in the national economy and is seen as catalyst in the creation of favourable systemic conditions for knowledge creation and an important actor within the National Innovation Systems framework and regional systems of innovation.

Engineering an appropriate welfare mix is the entrepreneurial challenge of state action. The outcome of successfully meeting this challenge is the development of an environment that harnesses and builds on the resources, including cultural capital, of the nation and mitigates the risks of the global ‘invisible hand’, yet allows its citizens access to the opportunities opened up by globalisation. The policy making and changing processes of the strategic state are thus geared toward enhancing the capacity of its economy and capabilities of its people through targeted intervention, within the context of the imperatives of the global age. It must also actively determine social policy and a social protection agenda that takes into account the demands and features of the global age. Social protection systems need to adjust to a new world of work with growing amounts of precarious non-standard work, non-linear employment and career paths and a need for lifelong learning. Under these

circumstances, the strategic state becomes the principal actor in laying the foundations for building a strong, socially inclusive economy within the globally connected world.

Conclusion

The new global era of capitalism has brought change to the contours in the landscape of state action, form and function. From an historical perspective, the ‘welfare state’ was an integral part of the previous era and its incompatibility in the context of the emergence of a new phase of capitalist development can therefore be questioned. “We need a new welfare state” (Esping-Andersen et al. 2002). Whatever the state forms that emerge and evolve in line with the demands of the global era, and the labels ascribed to different welfare mixes, however, they will undoubtedly be embedded within specific national traditions and policy legacies, political circumstances and party politics and ideological predilections.

Each country or area needs to design its ‘Own Way’. That is, it requires a specific model based on a proactive state that takes into consideration the nation’s development and social protection needs, within the overarching constraints and opportunities afforded by the global age. In designing such a model, the building blocks for a ‘new welfare architecture’ recommended by Esping-Andersen (2002) can be fruitfully used. Of particular value would be use of the methodology of a life-course framework, and the rethinking and framing of welfare as a social investment. The former, is a diagnostic methodology enabling an informed look into the future and which connects fragmented policy areas to capture the dynamics of citizens’ life chances since well-being conditions at one life cycle stage are often directly linked to other stages, e.g., poverty in old age is often a result of unsatisfactory employment outcomes during working years.

The latter, provides a perspective to allow escape from the commonly held notion that social outlays are unproductive current consumption without economic return. Rethinking social accounting practice, e.g., regarding a powerfully child oriented family policy as social investment, is a firm step forward for the creation of a sound, sustainable welfare architectural foundation.

Selected References

- Aspalter, C. (2001) *Identifying Variations of Conservative Social Policy in North East Asia: The Welfare State in Japan, South Korea and Mainland China*. Australian National University Graduate Program in Public Policy Discussion Paper, 81.
- Atkinson, A. (1995a) "Is the Welfare State Necessarily an Obstacle to Economic Growth?" *European Economic Review*, 39, 3, 723-730.
- Atkinson, A. (1995b) "The Welfare State and Economic Performance", *National Tax Journal*, 48, 2, 171-198.
- Audretsch, D. and Thurik, R. (1999) "Capitalism and Democracy in the 21st Century: From the Managed to the Entrepreneurial Economy", *Journal of Evolutionary Economics*, 10, 17-34.
- Auer, P. (2000) *Employment Revival in Europe: Labour Market Success in Austria, Denmark, Ireland and the Netherlands*. Geneva: International Labour Office.
- Barr, N. (1993) *The Economics of the Welfare State*. Second Edition. London: Weidenfeld and Nicholson.
- Barr, N. and Whynes, D. (1993) "Introductory Issues", in N. Barr and D. Whynes (Editors), *Current Issues in the Economics of Welfare*. Houndmills and London: Macmillan, 1-19.
- Beck, U. (1992) *Risk Society: Towards a New Modernity*. London: Sage Publications.
- Bowles, P. and Wagman, B. (1998) "Globalization and the Welfare State", *Eastern Economic Journal*, 23, 3, 317-336.
- Boyer, R. (1995) "Capital-Labour Relations in OECD Countries: From the Fordist Golden Age to Contrasted National Trajectories" in J. Schor and J. You (Editors), *Capital, the State and Labour: A Global Perspective*. Aldershot, UK: Edward Elgar and Tokyo: United Nations University Press, 18-69.
- Castles, F. (1985) *The Working Class and Welfare: Reflections on the Political Development of the Welfare State in Australia and New Zealand, 1890-1980*. Wellington: Allen and Unwin.
- de Bruin, A. (2003) "State Entrepreneurship", in A. de Bruin and A. Dupuis (Editors), *Entrepreneurship: New Perspectives in a Global Age*. Aldershot, UK: Ashgate, 148-168.
- Esping-Andersen, G. (1990) *The Three Worlds of Welfare Capitalism*. Cambridge: Polity Press.
- Esping-Andersen, G. (1999) *Social Foundations of Post-industrial Economies*. Oxford: Oxford University Press.
- Esping-Andersen, G. (2002) "Towards the Good Society Once Again", in G. Esping-Andersen, with D. Gaillie, A. Hemerijck, and J. Myles (Editors), *Why We Need a New Welfare State*. Oxford: Oxford University Press, 1-25.
- Esping-Andersen, G., with D. Gaillie, A. Hemerijck, and J. Myles. (2002) *Why We Need a New Welfare State*. Oxford: Oxford University Press.
- Fraser, D. (1973) *The Evolution of the British Welfare State*. London: Macmillan.
- Galbraith, J.K. (1992) *The Culture of Contentment*. New York: Houghton Mifflin.

- Hemerijck, A. (2002) "The Self-Transformation of the European Social Model(s)" in G. Esping-Andersen, with D. Gaillie, A. Hemerijck, and J. Myles (Editors), *Why We Need a New Welfare State*. Oxford: Oxford University Press, 173-213.
- Hicks, A. and L. Kenworthy. (2003) "Varieties of Welfare Capitalism", *Socio-Economic Review*, 1, 1, 27-61.
- Jessop, B. (1994), "The Transition to Post-Fordism and the Schumpeterian Workfare State", in R. Burrows and B. Loader (Editors), *Towards a Post-Fordist Welfare State?* London and New York: Routledge, 13-37.
- Johnson, N. (1987) *The Welfare State in Transition*. London: Harvester Wheatsheaf.
- Lindbeck, A.; P. Molander; T. Persson; O. Petersson; A. Sandmo; B. Swedenborg and N. Thygesen. (1994) *Turning Sweden Around*. Cambridge, MA: MIT Press.
- Mishra, R. (1999) *Globalization and the Welfare State*. Cheltenham, UK & Northampton, US: Edward Elgar.
- OECD (1981) *The Welfare State in Crisis*. Paris: OECD.
- Pfaller, A.; I. Gough. and G. Therborn. (1991) (Editors), *Can the Welfare State Compete?* Houndwills and London: Macmillan.
- Pfaller, A. with I. Gough. (1991) "The Competitiveness of Industrialised Welfare States: A Cross-country Survey" in A. Pfaller, I. Gough and G. Therborn (1991) (Editors), *Can the Welfare State Compete?* Houndmills and London: Macmillan.
- RCSS (1972) *Social Security in New Zealand*. Wellington: Government Printer.
- Sarfati, H. (2002) "Labour Market and Social Protection Policies: Linkages and Interactions", in H. Sarfati and G. Bonoli (Editors), *Labour Market and Social Protection Reforms in International Perspective: Parallel or Converging Tracks?* Aldershot: Ashgate, 11-57.
- Spicker, P. (2000) *The Welfare State: A General Theory*, Sage Publications, London.
- Stanfield, J.R. and B. McClintock. (1999) "Welfare State" in Phillip Anthony O'Hara (Editor), *Encyclopedia of Political Economy*. London and New York: Routledge.
- Taylor-Gooby, P.; H. Dean; M. Munro and G. Parker. (1999) "Risk and the Welfare State", *British Journal of Sociology*, 50, 2, 177-195.
- Titmuss, R. (1968) *Commitment to Welfare*. London: Allen & Unwin.

Anne de Bruin
 Colleg of Business
 Massey University,
 Aukland, New Zealand
 a.m.debruin@massey.ac.nz

Wetlands Governance

Firooza Pavri

Introduction

Wetlands refer to those unique ecosystems that lie on the cusp between the terrestrial and aquatic worlds. They are found in almost all regions of the world with the exception of Antarctica and have been subject to significant alternations by humans. Wetland governance generally involves both the national and international legal and institutional context that allows for the use and management of these ecosystems. All wetlands are characterized by the presence of standing water for at least part of the year. The presence of water creates unique soil conditions, which in turn support organisms and plants that have adapted to these conditions. The vast range of wetland habitats are influenced by locally specific climatic and soil conditions making a universally accepted scientific definition a challenge. Further complicating this task are the numerous common names applied to different types of wetlands and their variable use across geographic contexts. Marsh, swamp, bog, peatland, mangrove and vernal pool are perhaps the more often used terms, whereas fen, billabong, moor, wadden, sunderban and pantano reveal distinctive regional influences (Mitsch & Gooselink 2000).

Countries have generally phrased definitions to encapsulate the range of wetland habitats present, and policy frameworks particular, to their context. For instance, in 1979 the United States Fish and Wildlife Service adopted the following definition still used by wetland scientists in the United States today: "Wetlands are lands transitional between terrestrial and aquatic systems where the water table is usually at or near the surface or the land is covered by shallow water. Wetlands must have one or

more of the following three attributes: (1) at least periodically, the land supports predominantly hydrophytes; (2) the substrate is predominantly un-drained hydric soil; and (3) the substrate is nonsoil and is saturated with water or covered by shallow water at some time during the growing season of each year" (Cowardin 1979). This broad ranging definition covered many unique wetland characteristics and was flexible enough to accommodate both scientists and policy makers charged with wetland management. A less technical, yet equally all encompassing definition officially adopted by Canada included the following: "Land that is saturated with water long enough to promote wetland or aquatic processes as indicated by poorly drained soils, hydrophytic vegetation and various kinds of biological activity which are adapted to a wet environment" (Environment Canada 1986).

History and Conservation of Wetlands

Historically, wetlands have sustained human populations by acting as sources of food, water, fiber and shelter. Examples range from Sumer and early Egypt's cultivation of reed beds for fiber and building materials, to Chinese, Mayan and Aztec experimentation with wetland plant agriculture (Boule 1994). Boule's (1994) historical analysis also indicates that Mesopotamians valued the aesthetics of wetland landscapes and sought to replicate these by creating wetland gardens. Even so, these early agriculturalists recognized the economic value of drained wetlands as cropland, and the push for agricultural expansion led to wetland conversion across various parts of the world well into the 20th century. The industrial age only quickened this process as increased population pressures turned to heretofore untouched wetlands for development and profit. Policy initiatives from colonial India to North America pursued sustained wetland

conversion strategies, which were aided by the lack of scientific knowledge about the role of these ecosystems within the larger biosphere and their often erroneous representation as “wastelands.”

Beginning in the second half of the 20th century, however, numerous studies began documenting the valuable economic and ecological functions provided by wetland habitats at local, regional and global scales (Richardson 1994; Williams 1990; Wheeler et al. 1995). Scientists examined the complexity of wetland ecosystems and graphed feedback loops and decreased ecological productivity ensuing from wetland conversions to other land uses (Mitsch and Gooselink 2000). These studies also illustrated wetlands’ significant hydrological and biogeochemical contributions. It is now commonly accepted that wetlands improve water quality through toxin removal, they allow for the cycling of organic and inorganic nutrients through the ecosystem, they perform important roles in the nitrogen, sulfur, methane and carbon cycles, they accelerate groundwater recharge, and protect against erosion from coastal storms and flood events.

International Wetland Policy

The accumulation of evidence and emergence of the science of wetland ecology undoubtedly bolstered conservation and restoration legislation across nations of the world. As a further boost, the 1971 Convention on Wetlands of International Importance especially as Waterfowl Habitat, also known as the Ramsar Convention or the Convention on Wetlands, was signed in Ramsar, Iran. This intergovernmental treaty outlined cooperation between nation states in wetland habitat protection and conservation (US Fish and Wildlife Service 1993). While the original intent of the Convention was primarily to provide a habitat for water birds (as reflected by its name), the Convention

later broadened its scope to cover all aspects of wetland conservation. The Convention went into effect in 1975 after 7 countries ratified it and by 2006 there were 152 signatories to the treaty with 1596 protected wetland sites comprising 134.7 million hectares of habitat (Ramsar Convention 2006).

The Ramsar Convention does not have regulatory mechanisms in place to enforce the treaty. Rather, it relies on member nations keeping their wetland conservation and wise use obligations by developing national wetland policies in accordance with their overall natural resource planning strategies. Furthermore, it underscores the importance of establishing wetland nature reserves and member nations are to designate at least one wetland site within their borders as a wetland of “international importance” in terms of its “ecology, botany, zoology, limnology or hydrology.” More recently the Convention initiated National Ramsar Committees to provide support for the Treaty’s implementation at the national level (Ramsar Convention 2006). These committees are to involve a wide diversity of interested players including local government agencies, scientists, and members of non-governmental and community organizations. The Committees are charged with managing Ramsar sites, identifying new sites, providing expert knowledge, overseeing the implementation of new resolutions, and procuring funds for management through grants disbursed by the Ramsar Convention (Ramsar Convention 2006).

Cooperation between member countries is fostered through progress meetings. These include regular reviews of wetland sites on the list, and conferences focused on wetland habitat management, data collection and cooperation with other international conservation bodies. Financial contributions from member states support the Convention

and its administrative arm, the Ramsar Bureau, which is located in Gland, Switzerland at the headquarters of the World Conservation Union (Ramsar Convention 2006).

Criteria for identifying Ramsar wetlands of international importance fall into two categories. Group A Sites include areas containing representative, rare or unique wetlands, while Group B Sites include areas of international importance for conserving biological diversity. Representative examples of wetlands on the List include the more famous sites like the Pantanal in South America, the Everglades in North America, the Volga delta in Russia, the Okavango in southern Africa, the inland Niger delta in the Sahel, and the Sunderbans in South Asia. However, the majority of sites on the List include smaller and lesser known, yet equally valuable, wetlands important for their local and regional contributions (Ramsar Convention 2006).

National Wetland Policy in the Developed World

At the national level wetland governance has made significant strides in the later half of the 20th century. However, this progress in the protection and sustainable management of wetlands is far from uniform. Countries across the developed world have legislated and enforced stricter regulations for the conservation of these ecosystems while poorer nations, constrained by other societal needs, have been less inclined to promote legislation curbing the development or alteration of wetland environments.

The United States has been one of the early entrants into the realm of wetland protection and one of the countries with the greatest amount of legislative regulation for wetland habitats. Yet, this was not always the case. 19th century US policy toward wetlands was to impound, drain and alter them into

what were considered more economically productive land uses. This policy was promoted by the Swamp Land Acts of 1849, 1850, and 1860, which allowed for the drainage of wetlands and the expansion of agriculture and urban development (Wheeler et al 1995; Williams 1990). As Tzoumis (1998) reports, these and other policies were marked by the influence of pro-development philosophies and were precursors to dramatic change and acreage declines in wetland landscapes across the country. The 1960s, however, heralded a new era of environmental legislation and policy initiatives not necessarily limited to wetlands.

In the US early protective provisions for wetlands included the Clean Water Act of 1972 (Section 404) and its subsequent amendments. While Section 404 of the Clean Water Act never specified wetlands, it mandated that any dredging or filling activity within “waters of the US” to first obtain a permit from the US Army Corps of Engineers. Due to the lack of specificity on the part of the US Congress, the Army Corps of Engineers adopted a definition of “waters of the US” to include all navigable waters and wetlands by 1977. Meanwhile, early cases in front of the Courts used the Army Corps of Engineers’ expanded definition, thus laying down court precedent. The definitional confusion was based in part on the failure of Congress to address this issue adequately in 1972 or later in 1977 when it made amendments to the Clean Water Act (EPA 2005).

The permit system used by the Army Corps of Engineers necessitates that no dredging activity be permitted if alternatives exist. Furthermore, preliminary studies must be carried out at site to minimize the potential impact of those dredging and filling activities on wetlands, avoid the activity altogether where practical, or engage in mitigation activities for unavoidable impacts. Mitigation

includes providing compensation for unavoidable impacts through the restoration or creation of wetlands elsewhere. The Army Corps of Engineers cannot issue a permit if the wetland performs important biological and ecological functions, or when the ecological and economic costs outweighed the benefits proposed. The US Environmental Protection Agency (EPA) has the statutory authority to designate wetlands that are subject to permits and has the ability to veto the Army Corps of Engineers' decisions (EPA 2005).

Apart from the Clean Water Act, two other significant wetland policies also influenced wetland management in the US. They included two Presidential Executive Orders issued by President Jimmy Carter in 1977 to protect wetlands and riparian systems. The Protection of Wetlands executive order 11990 called for action by federal agencies to minimize the destruction, loss or degradation of wetlands and to preserve and enhance the values of these ecosystems, while the Floodplain Management executive order 11988 established similar federal policy for the protection of floodplains, requiring agencies to avoid activity in sensitive floodplain areas (Mitsch and Gooselink 2000). Federal agencies like the Environmental Protection Agency and the Natural Resources Conservation Service, among others, were compelled to review policies toward wetlands in light of these orders.

Finally, the No Net Loss Policy was an outcome of the 1987 National Wetlands Policy Forum convened by the Environmental Protection Agency to look at wetland management in the US. Among other issues, the Forum identified two goals for the nation's wetland policy. First, no overall net loss of the nation's remaining wetlands base, and second the restoration of wetlands where feasible (Mitsch and Gooselink 2000). This

did not necessarily argue for the cessation of all development, however, it made the strong case for remedial and restoration activities to mitigate the alteration of wetlands at another site.

In summary, US federal policy on wetlands is carried out under regulations related to land use and water quality rather than an overall national wetland law. As reported by Mitsch and Gooselink (2000), the federal government promotes the protection of wetlands through regulations (like Section 404 of the Clean Water Act), economic incentives and disincentives (for example, tax deductions for selling or donating wetlands to a qualified organization), and acquisitions (for example, by establishing national wildlife refuges). The enforcement of Section 404 includes issuing administrative compliance orders requiring violators to cease illegal activity, remove structures to restore sites, and assess civil penalties. For more egregious violations, the US Environmental Protection Agency and the Army Corps of Engineers exercise their criminal enforcement authority, which can lead to jail time and compensation costs (EPA 2005). At the state level, laws have been enacted to regulate wetland development activities and efforts taken by coastal states like Florida have reduced coastal wetland conversions. Counties and townships have also introduced zoning ordinances to regulate activities in wetland sites (EPA 2005). These and other measures have undoubtedly influence wetland conservation and as reported in recent studies indicate a marked decrease in the rate of wetland habitat loss across the US (Dahl, 2000).

Other noteworthy examples from the developed world include Canada's foray into wetland policy-making, which is marked with the distinction of being the first in the world to outline a federal wetland conservation policy (Government of Canada 1991). This

policy follows the recommendations of the Ramsar Convention on wise use and sustainable management. The Canadian government's 1992 Federal Policy on Wetland Conservation (FPWC) aims to achieve cooperation with the governments of its provinces, territories and private entities to maintain, enhance, rehabilitate, and secure wetlands of significance (Government of Canada 1991). It accomplishes these goals by promoting public awareness, developing exemplary wetland conservation practices, stimulating decision making based on scientific and technical factors, and enhancing cooperation between citizens, non-government organizations and governmental entities (Government of Canada 1991).

Differing significantly from the US approach, Canada has adopted a largely non-regulatory policy approach in that it relies primarily on the collective participation and cooperation of federal, provincial and territorial governments to undertake wetland conservation practices (Government of Canada 1991; Rubec 1994). While it may be difficult to gauge the long-term success of this policy given its recent adoption, the Canadian government has reported successes where the impacts of development have been limited through avoidance or adoption of mitigation strategies. Canada's wetland policy has also been example setting, as 25 other nations have now implemented similar non-regulatory policy measures. Given that Canada contains a significant 24% of total global wetland habitats and is a leading player in the Ramsar Convention with over 30 designated wetland sites, the results of this approach will be keenly studied (Rubec 1994). Furthermore, in the coming decades it will provide data for a comparative analysis between the pros and cons of regulatory and non-regulatory mechanisms in wetland policy making.

Rather than outline wetland-specific legislation, most European nations have opted to incorporate wetland protection within already existing laws on fishing, agriculture, water protection, or industrial development. And, since many western European states adopt a federal system, states or regions often play significant roles in not only enforcing existing federal laws, but also implementing regional wetland conservation regulations.

Germany presents one such example, wherein the German National Federal Conservation Act provides the umbrella policy specifically outlining that wetlands including fens, marshes, reed beds, meadows and other wetland ecosystems are protected, and development activity leading to their destruction or impairment is prohibited (Ramsar Germany Report 2006). Further, wetland protection is also folded into German legislation concerning pollution controls, and federal forestry, mining, and hunting laws. In so doing, it provides an additional measure of protection for wetland habitats. German states (or *Laender*) have the ability to further widen the scope of this protection by enacting supplementary policies that can be implemented regionally and where applicable. Alternatively, states also have the ability to provide permits for wetland conversions when it is in the public interest or where mitigation activities are employed; thus balancing development goals with conservation practices (Ramsar Germany Report 2006).

The United Kingdom, meanwhile, incorporates both site-specific and policy based mechanisms for wetland management and protection (Ramsar UK Report 2006). To that end, and as of 2006, there are over 164 Ramsar designated wetland sites across the UK and its territories, which are monitored and protected under various already existing federal laws including, most notably, the Wildlife and Countryside Act of 1981 and

other newer European Union directives on the conservation of natural habitats. A targeted approach adopted by the government identifies key habitat and species and necessitates the development and implementation of action plans for their protection. Overall, policy adopted across the UK is largely designed to incorporate Ramsar prescribed best use wetland practices, monitor significant sites, encourage land use policies that accommodate wetland habitats, and implement conservation strategies that are holistic (Ramsar UK Report 2006). All of these goals are supported by regulatory frameworks and implementation plans like the Estuary Plan, Shoreline Plan, or Water Management Plan. In the UK, government agencies have also made forging partnerships with private stakeholders an important aspect of their conservation goals. Beyond these efforts, non profit conservation organizations are active in the UK. One example includes the Wildfowl & Wetlands Trust of the UK, which has engaged in wetland habitat and species protection and restoration over the past 50 years. The Trust is now also supporting wetland conservation and education programs in countries across the globe (Wildflower and Wetlands Trust 2006).

While wetland management and conservation appears to be well established in Western Europe, it is in Eastern Europe, the CIS and Russia where management efforts and regulatory mechanisms still require greater articulation and enforcement. Most of this region inherited a legacy of environmental problems from policies adopted during the Cold War era. Pursuing industrial growth and development at the expense of environmental concerns was characteristic of the region for the second half of the 20th century. At present, the Ramsar organization is attempting to expand protection of important wetland sites while simultaneously urging the development of

greater country-based institutional and legal protections (Ramsar 2006). A cursory look at environmental and wetland legislation across this region suggests that it is piecemeal at best. Greater openness, however, and the documentation of environmental problems across the region will undoubtedly encourage calls for regulation and protection. Furthermore, ongoing technical and financial assistance from international conservation organizations have helped these countries initiate such discussions (Ramsar 2006).

Within the larger region, the West Siberian lowlands of Russia comprise some of the most extensive wetland ecosystems of the world. These lowlands range from east of the Ural mountains to the Yenisey river covering over two million sq km. Some estimates suggest that approximately half of this area is covered in wetlands (Solomeshch 2005). Peatlands, ranging between 1 to 5 m thick in some cases, occur across the landscape and were probably formed during the end of the last glacial period (Kremenetski et al. 2003). Despite the lack of consistent data, scientists believe these peatlands play an important role in global carbon sequestration. At present, 19 out of 35 currently designated Ramsar sites in Russia comprise peatlands and encompass approximately 9% of the total Russian Ramsar area (Ramsar Russian Federation Report 2006). While population pressures are not significant in this region, the exploration and extraction of minerals, timber, peat, oil and natural gas have all contributed to significant habitat alterations and declines in biodiversity. The former Soviet Union engaged in some site-specific wetland habitat protection in western Siberia as far back as the late 1950s (Solomeshch 2005). Based on this system, natural areas across Russia such as the *Zapovednik*, *National Park*, *Zakaznik*, and *Nature Monument* are categorized based on their size, the degree of protection they are

accorded, the flexibility of at-site management observed, and the extent to which sites are off-limits to development (Solomeshch 2005, p. 47). The Russian government has conferred such status to additional areas across the West Siberian lowland in an attempt to conserve these important resources. Meanwhile, monitoring key wetland ecosystem-health indicators is on-going at many sites (Ramsar Russian Federation Report 2006). These will provide a wealth of information on the efficacy of conservation practices. Moreover, the policing and enforcement of management plans ought to yield positive results in these protected areas. Even so, their effectiveness will only become apparent as long term monitoring efforts yield results.

National Wetland Policy in the Developing World

Wetland management issues across the developing world are marked by their own distinct set of challenges. Population and development pressures and the lack of institutional protection often exacerbate problems associated with the overuse of wetland resources across the developing world. Moreover, wetlands management often falls under the jurisdiction of various government ministries and in the absence of national wetland policies and protections, ministerial priorities and development pressures override wetland interests. For instance, in India, wetland management falls under the departments of agriculture, forestry, fisheries, revenue, and water resources, among others. Coastal mangroves fall under the control of the forest department, yet inland wetlands might be the responsibility of the revenue or agricultural department. Furthermore, subsidies provided for irrigation and fertilizers by the agricultural department, or development grants for aquaculture by the fisheries department could have a direct

negative consequence on wetlands also under their control (CES 2005).

Conversion to agricultural land, increased freshwater demand, coastal aquaculture, inland pisciculture, timber harvesting, and increased flows of agricultural and industrial effluents include just some factors responsible for the destruction of coastal and inland wetlands across the globe (Roygeri 1995; Whigham *et al.* 1993). The lack of awareness of wetlands' ecosystem benefits and their characterization in many geographic contexts as "wastelands" confound conservation efforts. Further complicating the task of habitat protection is when two or more countries share these ecosystems. Naturally, political relations between nations pose significant challenges as well. Major examples of ecosystems ranging over wide geographic extents include the Pantanal shared by Brazil, Bolivia and Paraguay, the Okavango shared by Botswana, Namibia, Angola and Zimbabwe, and the Sunderbans, which span across Bangladesh and India. In each of these cases, national policies alone cannot offer adequate protection. Joint agreements to share the benefits and protect these ecosystems often require the role of mediators to provide equitable solutions to complex management arrangements.

The Okavango river delta is one such case where cross-border animosities between Botswana and Namibia are exacerbated by conflicts over scarce water resources in this dry landscape. The designation of the Makgadikgadi Salt Pans (into which the Okavango river eventually drains) as a protected Ramsar site also complicates any joint protection efforts (IRN 2005). The Okavango wetland is a seasonally filled inland delta for the Okavango River, which has its headwaters in western Angola and makes its way through Namibia before disappearing in the midst of Botswana's Kalahari Desert. The river weaves a maze of

lagoons, channels and islands before disappearing into the desert. This landscape, and the rich soda deposits left behind as the water evaporates or dries into the desert, sustains a unique ecosystem famous for wildlife, including elephants and hundreds of bird species (Green Cross International 2005; IRN 2005). Both bilateral and multilateral agreements have played a role in the management of the Okavango river basin. One such tripartite agreement, the Permanent Water Commission on the Okavango reached by Namibia, Botswana and Angola has so far performed adequately as it seeks to equitably share water resources across borders (Green Cross International 2005). In such cases, governmental and non-governmental mediators like the non-profit Green Cross International play important roles in creating the atmosphere for fair solutions or providing technical and scientific assistance. In the case of the Okavango, the Global Environmental Facility has also funded projects and provided management assistance.

The Pantanal provides yet another example of the need for multilateral arrangements for wetland protection. The Pantanal is part of the Parana-Paraguay river basin (in Bolivia, Paraguay, and Brazil) and Mato Grosso (Brazil). It extends over approximately 140,000 sq miles and comprises riverine, palustrine and lacustrine wetlands ecosystems (Banks 1991). According to Wais and Roth-Nelson (1994), the Parana river is one of the most intensively developed rivers in South America and is the main focus for development in the coming decades. Twenty three hydroelectric projects are either being planned or have already been built across the Parana river systems to 'tame' the rivers, provide an internal transportation corridor and boost development in the interior of these countries. This, along with increased population and agricultural and industrial development pressures promises to change

the Pantanal's habitat significantly (Junk 1993; Swartz 2000). Cooperation between countries has thus far largely focused on the development of the Parana basin. Conservation and management of wetland resources across the region have been less of a concern during multi-party talks. As with other developing contexts, the responsibility for wetland management is generally the concern of various government ministries and rarely comprises a uniform policy.

Mangroves or subtropical and tropical coastal wetlands have, in particular, witnessed significant decline over the past century (Richards 1990). These unique wetland habitats have been shown to provide integral ecological functions, including forming the basis of complex marine food chains. Today, these habitats are disappearing at an accelerating rate all across South and Southeast Asia, instead making way for shrimp aquaculture and other industrial farming practices (Baird and Quarto 1994). For instance, Thailand has lost almost half of its mangrove forests since 1960 (EII 2005). Governments are now recognizing the costs associated with this rapid destruction and in certain cases have taken protective measures. The Indian and Bangladeshi governments have made attempts to carefully manage shared mangrove forests covering an area of approximately 10,000 sq km along the Ganges delta. The Sunderban Biosphere Reserve and the mangrove eco-park in Jharkhali are both recent efforts to protect this ecosystem (CEERA India 2005). Yet, increasing populations within the region have led to significant development pressures on these resources and it remains to be seen if these efforts will sustain themselves in the long run.

21st Century Monitoring and Management

This review suggests that wetland policy across the globe is far from uniform. While

the second half of the 20th century has seen significant steps toward wetland conservation and protection, far more needs to be done to jumpstart these efforts across parts of the developing world, Eastern Europe, the CIS and Russia. There appear to be two emerging philosophies that guide wetland management. On the one hand, the regulatory approach adopted by the US and other nations focuses on laws, enforcement mechanisms, and penalties to protect these habitats. Canada leads a second, largely non-regulatory, approach designed to build partnerships and persuade the public to support broad wetland conservation policies. It remains to be seen which of these approaches is more successful.

Human-induced threats to wetlands are only becoming more acute. Ever increasing demands for water to meet agricultural, industrial and domestic needs add severe strains to an already stretched resource. The implications of climate change, increased frequency of severe events, and rising sea levels will have ramifications on natural ecosystems, wetlands included. In the absence of effective conservation regulations, increasing land use pressures to meet food production will result in wetland habitat alterations as has already seen to be the case across the world. As a recent Ramsar report suggests, the daunting challenge for nations will be to craft legislation and build conservation strategies ensuring sustainable wetland use in the context of these global threats (Ramsar Strategic Plan 2006).

Satellite imagery and aerial photographs have provided unique and systematic views of the earth's ecosystems. Due to the general difficulties of traversing wetland habitats by foot, aerial photography and now satellite imagery technologies are routinely used to monitor wetland change and document the impact of human activities on these vital resources (Lyon 2001). Over the past few decades these imagery have provided natural

resource managers with important data about the efficacy of their management efforts. In some cases, they have also persuaded policy makers to endorse more stringent management strategies. Regardless, these data not only provide the scientific community with a better understanding of the functioning and health of these ecosystems, but also indicate that a great deal needs to be done to document and monitor these habitats. Such studies would not only improve our understanding of these unique landscapes and their functioning, but perhaps also provide convincing evidence for their future conservation.

Selected References

- Banks, V. (1991) *The Pantanal: Brazil's Forgotten Wilderness*. San Francisco: Sierra Club Books,
- Baird, Ian and Alfredo Quarto. (1994) *The Environmental and Social Costs of Developing Coastal Shrimp Aquaculture in Asia*. San Francisco: Earth Island Institute.
- Boule, M. E. (1994) "An Early History of Wetland Ecology", in W.J. Mitsch (Editor), *Global Wetlands: Old World and New*. Amsterdam: Elsevier 57-74.
- CEERA India. (2005) *Environmental Updates: Sunderbans*. www.nls.ac.in/CEERA/ceerafeb04/html/index.htm
- CES. (Centre for Ecological Sciences) (2005) *Report on Wetland Management in India*. Bangalore: CES.
- Cowardin, L. (1979) *Classification of wetlands and deepwater habitats of the United States*. Washington DC: Fish and Wildlife Service, US Dept. of the Interior.
- Dahl, Thomas. (2000) *Status and Trends of Wetlands in the Conterminous US 1986-1997*. Washington DC: Fish and Wildlife Service, US Dept. of the Interior, 1-82.

- Dugan, P.J. (1994) "Wetlands in the 21st Century: The Challenge to Conservation Science", in W.J. Mitsch (Editor), *Global Wetlands: Old World and New*. Amsterdam: Elsevier, 75-87.
- EII. (Earth Island Institute) (2005) *Mangrove Action Project*. www.earthisland.org/map
- Environment Canada. (1986) *Wetlands in Canada: A Valuable Resource*. Ottawa, Ontario: Lands Directorate, Fact Sheet Number 86-4, 1-8.
- Environment Canada. (2005) www.ec.gc.ca/water/en/nature/wetlan/e_wetlan.htm
- EPA. (US Environmental Protection Agency. (2005) Section 404 of the Clean Water Act: An Overview. www.epa.gov/owow/wetlands
- Government of Canada. (1991) *The Federal Policy on Wetland Conservation*. Ottawa, Ontario: Environment Canada, 1-14.
- Green Cross International. (2005) *The Okavango River Basin*. www.gci.ch/index.htm#
- IRN (International Rivers Network) (2005) *IRN's Okavango Campaign*. www.irn.org/programs/okavango
- Junk, Wolfgang. (1993) "Wetlands of Tropical South America" in Dennis D. Whigham, Dagmar Dykyjova; and S. Hejny (Editors), *Wetlands of the World: Inventory, Ecology and Management*. Volume 1. Dordrecht: Kluwer, 679-739.
- Kremenetski, K.; A. Velichko; O. Borisova; G. Macdonald; L. Smith; K. Frey and L. Orlova. (2003) "Peatlands of the Western Siberian Lowlands: Current Knowledge on Zonation, Carbon Content and Late Quaternary History", *XVI Inqua Congress*. Reno, NE: Geological Society of America.
- Lyon, J. (2001) *Wetland Landscape Characterization: Techniques and Applications for GIS, Mapping, Remote Sensing and Image Analysis*. Chelsea: Ann Arbor Press.
- Mitsch, W.J. and J.G. Gooselink. (2000) *Wetlands*. Third Edition. New York: Wiley and Sons.
- Ramsar Convention. (2006) www.ramsar.org
- Ramsar Germany Report. (2006) www.ramsar.org/cop9/cop9_nr_germany.pdf
- Ramsar Russian Federation Report. (2006) National Report of the Russian Federation for the Period 2003-2005. www.ramsar.org/cop9/cop9_nr_russia.pdf
- Ramsar Strategic Plan. (2006) Ramsar Strategic Plan, 2003-2008. www.ramsar.org/key_strat_plan_2003_e.htm
- Ramsar United Kingdom Report. (2006) *National Planning Tool for Implementation of Ramsar Convention on Wetlands*. www.ramsar.org/cop9/cop9_nr_uk.pdf
- Richards, J.F. (1990) "Agricultural Impacts in Tropical Wetlands: Rice Paddies for Mangroves in South and Southeast Asia", in Michael Williams (Editor), *Wetlands: A Threatened Landscape*. Oxford: Basil Blackwell.
- Richardson, C.J. (1994) "Ecological Functions and Human Values in Wetlands: A Framework for Assessing Forestry Impacts", *Wetlands*, 14, 1, 1-9.
- Roygeri, Henri. (1995) *Sustainable Management: Guiding Principles and Practical Approaches in Tropical Freshwater Wetlands*. Dordrecht: Kluwer.
- Rubec, C.D.A. (1994) "Canada's Federal Policy on Wetland Conservation: A Global Model", in W.J. Mitsch (Editor), *Global Wetlands: Old World and New*. Amsterdam: Elsevier, 909-917.
- Solomeshch, A.I. (2005) "The West Siberian Lowland", in L.H. Fraser and P.H. Keddy (eds), *The World's Largest Wetlands: Ecology and Conservation*. Cambridge: Cambridge University Press, 11-62.
- Swarts, Frederick A. (2000) (Editor) *The Pantanal of Brazil, Bolivia and Paraguay*. Gouldsboro, PA: Hudson MacArthur Publishers.

- Tzoumis, K.A. (1998) "Wetland policymaking in the U.S. Congress from 1789 to 1995", *Wetlands*, 18, 3, 447-459.
- U.S. Fish and Wildlife Service. (1993) *Wetlands of International Importance: United States Participation in the "Ramsar"*. Washington DC: Convention, Department of the Interior.
- Wais, I.R., and W. Roth-Nelson. (1994) "Management Strategy for a Large South American Floodplain Wetlands System: The Parana-Paraguay Basin", in W.J. Mitsch (Editor), *Global Wetlands: Old World and New*. Amsterdam: Elsevier, 713-723.
- Wheeler, B., S.C. Shaw, W.J. Foijt, and R.A. Robertson. (1995) (Editors) *Restoration of Temperate Wetlands*. Chichester: Wiley & Sons.
- Whigham, Dennis; D. Dykyjova, and S. Hejny. (1993) *Wetlands of the World: Inventory, Ecology and Management*. Volume 1. Dordrecht: Kluwer.
- Wildflower and Wetlands Trust. (2006): www.wwt.org.uk/visit/wetlandcentre/default.asp
- Williams, M. (1990) (Editor) *Wetlands: A Threatened Landscape*. Oxford: Blackwell.

Websites

- Ramsar Convention on Wetlands. www.ramsar.org
- International Rivers. www.irn.org
- Waterland Research Institute. www.pantanal.org

Firooza Pavri
Department of Geography-Anthropology
University of Southern Maine
Gorham, Maine, USA
fpavri@usm.maine.edu

Worker Cooperatives and Participatory Enterprises

Roger Ashton McCain

Introduction

A worker cooperative is a democratic enterprise in which "control rights follow from membership in the firm's workforce and ownership by itself confers no decision-making rights" (Bonin, Jones & Putterman 1993:1307). Participatory firms are enterprises in which workers may participate in management decisions or in profits or both to an extent that (intentionally or not) approximates a worker cooperative. Enterprises of these kinds pose two issues of governance. First, as small "republics in the workshop", cooperative and participatory enterprises face issues of their internal governance. Second, the establishment or public support of cooperative and participatory enterprises may promote the aims of public policy in direct or indirect ways. These issues are somewhat interdependent and this essay will explore both.

Cooperation, Ownership, and Control

In discussing cooperative and participatory firms, we will need first to distinguish among several closely related but distinct forms: employee ownership, workers' cooperation, workers' control and participation in decisions, and profit sharing, among others. In the words of Dow and Putterman (2000) "There is far too much packed into the ordinary concept of 'firm ownership' for it to serve as a useful primitive ..." and, in another sense, there is too little. A worker cooperative is an association of labor suppliers that operates according to the cooperative principles as set out by the International Cooperative Alliance (1995). Among these principles, the first four, voluntary and open

membership; democratic control by the members; member participation in the surplus; and autonomy, together define an ideal for the organization of production. The remaining three, information, cooperation among cooperatives, and concern for the community, extend the ideal to envision a whole society consistent with the first four. This ideal makes membership a first principle, and treats ownership as a secondary matter, a means to the end of realizing the cooperative principles. Worker cooperative enterprise offers a clear third alternative to either capitalist or government organization of production.

Because of their democratic character, worker cooperatives and participatory enterprises can be seen as carriers of a value beyond their economic contribution. Let us define democracy as the responsibility of those who exercise authority to those over whom the power is exercised. Democracy in political institutions, however imperfect, fits this simple definition. The significance of participatory enterprise and worker cooperation then is the fact that they extend democracy into the firm, making those who exercise authority in the workplace responsible, not to some distant parliament or shareholders, but to those over whom they exercise authority.

This has been widely recognized. Vanek (1971) says that worker cooperation has a special dimension beyond its economic advantages. Lutz and Lux (1979) and Lutz (1997) stress the value of cooperative enterprise in the light of their humanistic economics. Ellerman, (1990) drawing on the ideas of Locke, argues that the workers in an enterprise have an inherent right to ownership of the firm, having "mixed their labor" with it. This arises from Ellerman's Labor Theory of Property. This can be further supported from a surprising source. Nozick, (1974) in an argument for untrammelled markets, takes

it as given that a social arrangement (such as property) is just to the extent that it arises by just steps (such as mutually voluntary exchange) from a just starting point. But (a difficulty that Nozick does not raise) because capital is fungible, any sum of actual capital may be derived at once from just and unjust circumstances. Can any owner of capital really say that his wealth is not traceable in the least to piracy and slave-raiding? On the other hand, the one resource that a person may be said without doubt to have a just property in is her or his own labor. Thus, we can conclude with Ellerman (but by a quite different route) that only property founded in labor is just, and consequently that conventional ownership of large firms can never be equally undoubtedly just.

These non-economic values are relevant to governance, which should reflect a wide range of values. However, much of the discussion of worker cooperation and participation since the 1960's have focused on the economic functioning and advantages of cooperation, in the spirit of a worst-case analysis, and that perspective will largely be adopted for the remainder of this essay.

Among real enterprises, profit sharing, limited worker participation in management decisions, and some degree of worker ownership can be found both in enterprises intended to realize the cooperative principles and in others, state and capitalist enterprises, that do not. Each of these three characteristics may be found in the absence of the other two. This essay will be limited to enterprises founded on the cooperative principles or which involve substantial "employee participation" in management decisions, profits, or both. "Conventional firms" are other capitalist enterprises. Employee stock ownership plans—"ESOPs"—are usually organized so as to minimize worker participation in management decisions. (Kelso & Hetter 1967, Kelso & Adler 1958,

Blasi & Kruse 1991, Blasi, Conte and Kruse 1996) Thus, worker ownership based on these plans are not per se within the scope of this essay. Even so, "actual [worker cooperatives] are in fact quite heterogenous." (Bonin, Jones and Putterman 1993:1301) The most widely studied examples, the complex at Mondragon, the plywood cooperatives of the U.S. Pacific Northwest and the Italian and French cooperative sectors, can all employ nonmember labor, and require members to buy membership; participation in decisions by members may in practice be slight, and open membership may be more or less restricted. It seems that these enterprises at best approximate the ideal, and approximate it in different ways.

There is a large body of economic theory of cooperative enterprise (e.g. Ward 1958; Vanek 1970; McCain 1977; Bonin 1981), but it generally is based on some aspects of the ideal rather than on empirical studies (Bonin, Jones and Putterman 1993). This focus may not be a bad thing. To the extent that the actual enterprises represent different attempts to realize the ideal, and may be refined over time or may be abandoned (Gunn 1992) the theoretical studies based on the ideal may be more useful for some purposes. Practice can only benefit from the dialog among economic theory, the ideals expressed in the cooperative principles, and empirical studies of real enterprises founded to some considerable extent on those principles.

Economic Theory of Worker Cooperatives

Although discussion of worker cooperatives and the founding and operation of worker cooperatives were both extensive in the nineteenth century (Mill 1909/1987; Jones 1898/1968; Ely 1901), the modern economic theory of worker cooperatives begins with a paper of Ward, an attempt to characterize an abstraction of the economic system then established in Yugoslavia. In a bit of

geographic wordplay, Ward placed his worker cooperative economic system in Illyria, and the term "Illyrian" has since been used for the model of the worker cooperative that Ward explored. Ward adopted the production-function approach from neoclassical economics, and paralleled neoclassical economics by assuming that the decisions of the worker cooperative would be made in such a way as to maximize net income per worker. The production function approach treats an abstract and undifferentiated labor, L , and an abstract and undifferentiated capital good, K , as inputs and assumes that the maximum output that the enterprise can produce, Q , is determined by a function $Q=f(L,K)$. The labor input, L , is identified with the membership of the cooperative. The "production function", $f(\dots)$, is thought of as an expression of a technology that is equally available to any enterprise that may be established. In practice, it is necessary to assume that the "production function" has certain regular mathematical properties, such as continuity, smoothness, and "diminishing returns to a variable input."

The Illyrian approach also adopts the assumptions used in neoclassical economics to represent "perfect" price competition, i.e. that output is sold at a parametric price p and "capital" can be obtained at a parametric rental price or "rate of return" r . Income per member (Y) then is

$$Y = [pf(L,K) - rK]/L.$$

Ward then applies the methods of nonlinear programming to determine the values of L and K that maximize this income per member expression and the methods of vector calculus to explore the "comparative statics" of this maximization problem, contrasting these with the results for a hypothetical "capitalist" firm characterized by the same "production function." One striking result is that as the price, p , rises (under circumstances in which the "capitalist" firm would receive a profit

above the opportunity cost of capital) the membership and output of the Illyrian firm decline. Since profits over the opportunity cost of capital and a fixed number of firms are conditions that define the "short run" in neoclassical economics, this is expressed by saying that the Illyrian firm has a "backward sloping supply curve in the short run."

Much of the subsequent literature, both theoretical and empirical, has been organized around criticisms of Ward's Illyrian model. Early empirical studies (Berman 1967; Bellas 1972; Bernstein 1976; Jones & Backus 1977; Cable & Fitzroy 1980) established the following "stylized facts" about cooperative and participatory firms: 1) There is little or no evidence of the backward-sloping supply curve; 2) worker cooperatives tend to use less capital per worker than comparable conventional enterprises, as nearly as the neoclassical construct of undifferentiated capital can be approximated; 3) labor productivity is no less, and often is greater, in cooperative than conventional enterprises 4) both profit sharing and labor participation in management have favorable impacts on labor productivity, and they are interactive, tending to reinforce one another; and 5) there are important contradictions between Ward's assumptions and the organization of real worker cooperatives, at least in capitalist countries. (Jones 1985 establishes a similar contrast between cooperative and state enterprises in Poland.) Although this is not widely noted, 2) and 3) together imply that worker-cooperatives have the advantage over conventional firms in terms of factor-neutral productivity, and this generally contradicts Ward's assumption that the worker cooperative and the capitalist firm would have access to the same productive technology. Early theoretical work focused on the realism of the backward-sloping supply curves (Robinson 1967; Domar 1966) and on the realism of the assumptions of equal access

to capital and technology (Carson 1977; Jensen & Meckling 1979; Furubotn & Pejovich 1970; Furubotn 1976), although these critics assume that the conventional firm would have the productivity advantage.

One thing on which virtually all theoretical critics agree in rejecting, that the empirical studies reject, and that disagrees with many anecdotal accounts, is Ward's assumption that coops could have unlimited access to investment capital at a given "competitive" rate of return or rental cost. The empirical findings that coops typically use less capital per worker than comparable conventional enterprises is consistent with the observation of theorists that the "rented capital" model is unrealistic (Furubotn 1980). There could be at least two reasons why worker cooperatives have less access to capital than comparable conventional enterprises. First, bankers may not wish to lend to cooperatives simply because the cooperatives are an unfamiliar form. Second, cooperatives may be liquidity constrained.

To say that a potential borrower is liquidity constrained is to say that the liquidity available to the borrower through borrowing is constrained by the borrower's lack of assets to pledge as collateral or to assure that the lender can recover a substantial part of the amount due in the case of bankruptcy or default. Liquidity constraint could limit cooperatives more stringently than other enterprises for two reasons: first, because the founders of cooperatives are, on the whole, poorer than the founders of conventional firms and so have less assets to offer as collateral and assurance, and second, returning to the first reason, because bankers demand more collateral of an unfamiliar form of enterprise or one in which ownership is not the defining principle. On this assumption, there have been proposals for new kinds of financial instruments that might ameliorate this problem without sacrificing the

cooperative principles. (Vanek 1977; McCain 1977; Gui 1985; Major 1996). However, this approach remains untried.

Now, liquidity constraint presumably also applies to conventional firms, but generally plays no role either in modeling conventional firms for comparison with cooperative and participatory firms nor in economic theory generally. Liquidity constraint is closely associated with business bankruptcy (Hellwig 1981) and both are largely ignored in the neoclassical economics literature on the supposition that they are complexities that do not affect the theory in important ways. However, the limited literature on bankruptcy and liquidity constraint strongly suggest that this is not true. In a world of liquidity constraint, different firms may face different marginal rates of return to capital and these may rise with the amount borrowed (Kalecki 1937; Baumol 1953; Stiglitz 1969), consumers depart from the life-cycle rational consumption plan (Flavin 1981), and bankruptcy rules affect the allocation of real resources (Eichberger 1989; Hart 2000), raising the question whether the consensus neoclassical theory, which ignores all this, can be at all reliable as a guide to any real world market economy. Thus, while the consensus view that worker cooperatives are liquidity constrained seems correct, we should be very cautious about comparisons based on a view of conventional firms that supposes that they are not.

There is far less consensus in theory about the implications of worker cooperatives for effort and labor productivity. Critics from the property rights school routinely assume that effort and labor productivity will be lower in worker cooperatives than in conventional firms. (Furubotn & Pejovich 1970; Furubotn 1976, 1980; Carson 1977, Jensen & Meckling 1979) While their methods generally are mathematically informal (and therefore inconclusive) the line of reasoning here is

clear and easily put into mathematically rigorous form. First, they reject as mistaken the characterization of the labor input as undifferentiated. Instead, they propose that labor time and effort are to some extent independent. The "production function" might be $Q=f(EL,K)$ rather than $Q=f(L,K)$, where L is labor time supplied by the members and E is the average effort level supplied by the members. Labor productivity is supposed to be proportionate to effort in this formulation. (McCain 1980, seems to be alone in proposing explicitly that participatory enterprises may have a different effort-productivity tradeoff than conventional firms.) If we treat effort supply as a noncooperative game and assume that worker payoffs are increasing in Q (for which the objective function $[pf(L,K)-rK]/L$ is a particular case) then effort supply is a social dilemma. An efficient effort supply E^* can be defined which is Pareto-efficient from the members' point of view, but a lower effort supply will be the dominant strategy for each worker. As a result, unless the work group constitutes a cooperative coalition and enforces a high effort level, effort and labor productivity will be inefficiently low. The property-rights school assert that conventional firms will attain a cooperative allocation while worker-cooperatives will remain at the noncooperative equilibrium.

Put in this way—identifying cooperatives with a noncooperative equilibrium—this may seem odd, but there is no logical failure in the property rights theory, since cooperative (in the sense of the cooperative principles) and cooperative (in the game theoretic sense) simply are two different terms. However, a Platonic idealist might see a deeper inconsistency in it, taking cooperative (in the sense of the cooperative principles) and cooperative (in the game theoretic sense) as two imperfect expressions of a common underlying ideal of cooperative action. On the

other side, this noncooperative game model of effort and productivity poses a problem when it is applied to any hierarchical firm. To realize an efficient effort commitment, a group of workers will have to be supervised, but what incentive has the supervisor to supervise effectively? *Qui custodiet custodes?* The supervisor will have to have a supervisor, and so on; but the regression has to stop somewhere. Alchian argues that a proprietary firm is an ideal in that the regression stops at the proprietor, who receives the residual income and so has incentives to supervise those under his direction efficiently. This view certainly poses some difficult issues for a state-organized economy, in which the ultimate supervisor is the distant state (however democratic) and this may mean that supervision on the shop floor is ineffective. It also poses difficulties for corporate firms, in which the ultimate supervisor is the mass of anonymous shareholders. For a worker cooperative, however, there seems to be no good reason why the coop would not hire a supervisor, empower that person to enforce a high (cooperative in the game theory sense) effort level, and supervise him from the floor. As residual claimants, they would have the incentive to supervise the supervisor efficiently, no less than the proprietor than in a proprietary firm. (Craig & Pencavel 1995) This idealistic line of thinking leads us to expect that productivity would be high in proprietary and worker cooperative enterprises, lower in corporate enterprises, and lowest in state enterprises.

Moreover, recent experimental evidence of the importance of reciprocity in human motivations (McCabe, Rassenti & Smith 1996; Fehr & Fischbacher 2004) reinforce this idea. Workers acting with reciprocity would reward one another's higher efforts and punish one another's slacking, performing a kind of mutual supervision. This is consistent with observations that cooperatives commit

less resources to supervision, relying instead on a resource of mutual supervision that is not available to capitalist or state enterprises. Thus there is ample theoretical reason to expect that worker cooperatives will have an overall productivity advantage over other enterprise forms, as in McCain (2007).

This view is consistent with the empirical evidence. Early studies pointing in this direction have been mentioned. Doucouliagos (1995) did a meta-analysis of a number of econometric studies of the impacts of democratic decision-making, worker ownership, profit-sharing and collective ownership on productivity in worker cooperatives and participatory capitalist firms. He found that the first three tend to have positive influences on productivity, and the influence is stronger in the case of worker cooperatives than in the case of participatory capitalist firms. While there is no direct comparison of worker cooperatives with conventional firms, the tendencies to democratic decision-making, profit-sharing and worker ownership are more intensely developed in worker cooperatives, an overall advantage for this form is consistent with Doucouliagos' evidence. The fact that worker cooperatives often arise through the conversion and rescue of enterprises that are not viable as conventional firms (Spear and Thomas 1997, McCain 1999) also poses a paradox for any economic theory that does not recognize the efficiency advantages of cooperation.

Cases: USA

The study of worker-owned plywood companies in the northwestern United States, in particular, points in this direction. In these cooperatives, membership is bought at a price that fluctuates in the market, members are given first consideration for jobs that are open, but nonmember labor may be hired; all workers are paid the same wage; and the

supervisor is hired by a board of worker-members (Craig & Pencavel 1993,1994). Berman (1967), Bellas (1972) and Bernstein (1976) note tendencies toward mutual supervision on the part of the workers and strong supervisory powers available to the executive, a tendency for the coops to employ fewer supervisors per worker than conventional firms do, and tendencies to higher productivity of labor in the worker-owned plants than in comparable non-worker-owned plants. Craig and Pencavel (1995) find evidence that coops are more efficient than conventional firms by 6 to 14 percent, again consistently with the idealist view, a difference that is largely offset by their greater difficulty in raising capital. They also find that the cooperative firms respond quite differently in response to changing prices of inputs and outputs. The elasticity of supply is only slightly less for coops than for other types, and there is no backward sloping supply curve. This is also consistent with earlier studies of the plywood cooperatives.

Among the findings for the plywood cooperatives are that, in times of declining demand, they tend to reduce wages rather than laying off workers. (Craig & Pencavel, 1992) While this is evidence against a key conclusion of Ward's Illyrian model, it is probably more important for its implications for macroeconomic dynamics. This tendency suggests that an economy consisting largely of worker cooperatives would be less likely to experience output fluctuations over business cycles than is a capitalist economy. This tendency does not seem to have been studied for worker cooperative groups in other countries, however.

Cases: Spain

Smith (2003a:181) "argues that cooperatives may benefit ... [from] network externalities or complementarities of organizational type" in a region with relatively high frequency of

worker cooperatives. This “suggests that even if barriers to entry are overcome and a coop is established it may not survive, not because of intrinsic inefficiencies, but simply because of the lack of other cooperative entry, and to some extent also because of a lack of coordination among coops that do enter the market”, (p. 185). Moreover, he argues, the existence of a cooperative league may internalize some of these externalities. He uses (2003b) case studies of the Mondragon Cooperative Corporation in the Basque region in Spain and La Lega Nazionale delle Cooperative e Mutue (The National League of Cooperative and Mutual Societies), in Italy as illustrations of the thesis. The network externalities lead to a multiple-equilibrium model (Smith 2003a) in which one market equilibrium has little or no cooperative presence while other equilibria may have substantial presence of cooperatives and/or a League of Cooperatives. This important theoretical advance is consistent with the clustering of successful cooperatives in Spain, Northern Italy and other areas and the rarity of cooperatives in many other areas, and has important governance implications.

The Mondragon complex is an often-studied instance of worker cooperation that is commonly offered as a model. (Thomas & Logan 1982, Whyte & Whyte 1988, Lutz 1997, Birchall 1997:98-103) The Mondragón Cooperative Corporation (MCC) arose from a school and a complex of worker cooperatives established around 1960 through the efforts of a popular parish priest, José María Arizmendiarieta, known at Mondragón as Don José María, and usually referred to in the literature as Arizmendi. (Mondragon Corporacion Cooperativa 2001) Worker cooperatives in the Mondragon group are required adhere to a set of principles very much like those of the International Cooperative Alliance and to set aside a certain part of their revenues for capital

formation and job creation. Membership is open to all employees based on a uniform share purchase and a probationary period. In the early period, Mondragon cooperatives had the financial support of the mutual bank, Caja Laboral Popular, and today benefit from the credit, insurance and mutual-support activities of the MCC, which is itself organized as a democratic and cooperative joint body. The Mondragon cooperative network employs over 45,000, of whom more than half are members or probationary members. A major, diversified conglomerate enterprise, MCC struggles to globalize its operations without losing its commitment to cooperation. (Errasti, Heras & Bakaikoa 2003; Bakaikoa; Errasti & Begiristain 2004). The Mondragon experiment has been successfully imitated in other Spanish regions, such as Valencia. Valencia has had fast growth of worker cooperatives in recent decades (Chaves 1998). Development of worker cooperatives and participatory firms with worker ownership has been important also in Andalusia (Romero & Perez 2003) and Catalonia (Spear & Thomas 1997).

Cases: Italy and France

The National League of Cooperative and Mutual Societies, known in Italy as Legacoop, or simply La Lega, founded in 1886, defines itself as a league of autonomous cooperative enterprises. The larger part of La Lega are agricultural cooperatives, consumer cooperatives, housing cooperatives and others, but La Lega incorporates about 5000 worker cooperatives. The principles governing these worker cooperatives are similar to those of the International Cooperative Alliance, of which La Lega is a section, and Mondragon. A study of this sector concludes that it is characterized by higher productivity and better labor relations than comparable Italian conventional enterprises, consistently with other empirical

work (Bartlett et al. 1992). The rules of La Lega permit secession from the network, but such secession is uncommon.

The French Sociétés Coopératives Ouvrières de Production (SCOP), although smaller and limited to worker cooperatives, is a group with similar roots to La Lega. (Defourney, Estrin & Jones 1985; Perotin 1987; Defourney 1992; Estrin & Jones 1992; Bataille-Chetodel & Huntinger 2004) Empirical studies of this group have been largely consistent with findings on worker cooperatives in other countries, with some tendency toward higher labor productivity offset in some cases by lesser access to capital. It is also observed (Perotin 1987) that the cooperatives have a higher overall survival rate than comparable conventional firms. This is of interest as it contradicts theoretical predictions, both from the Illyrian and the property rights model, that worker cooperatives would display "self-extinction tendencies", but this has not been explicitly studied for groups in most other countries.

These French and Italian worker cooperatives are subject to a requirement that membership be open to all employees, subject to payment of a uniform nominal buy-in fee. While most theoretical work has not allowed for such openness, Kamshad (1997) has explored a model of equilibrium in a cooperative with open membership and finds that it has quite different implications. Kamshad's model is broadly but not precisely based on the French cooperatives regulations. The requirement that employees may become members at will, by buying one share at a fixed nominal price, assures that the cooperative cannot degenerate into a conventional firm (as an Illyrian firm might and some cooperatives in other countries have) by reducing the number of members through attrition and so increasing the incomes of the dwindling number of members. The conversion from employees to

members also assures a supply of capital at a market rate so long as there are nonmember employees. (Kamshad does not allow for liquidity constraint either of the cooperative or its individual members). Kamshad's model predicts both lower rates of entry and exit for worker cooperatives than for conventional (or Illyrian) firms, *ceteris paribus*. Kamshad's prediction that worker cooperatives under this regulatory regime would tend to be long-lived agrees with Perotin's empirical findings on the lower exit rate of French worker cooperatives than of comparable conventional firms.

Cases: The Rest of the World

The countries of the European western Mediterranean are today the site of the most intensive development of worker cooperatives. (Birchall 1997:97; Spear & Thomas 1998; Chaves 1998) From the 1950's through the 1980's, Yugoslavia had an economic system based in principle on workers' control of enterprises. (Prasnikar & Svejnar 1988; Nishizumu & Page 1982) The extent to which this was fully realized has inevitably been a topic of debate (Birchall 1997:119) and in any case was a moving target as the non-democratic Yugoslav government shifted its regulations. This discussion will have to be beyond the scope of this essay. Some remnants of this system have survived "privatization" in at least some post-Yugoslav republics. (Prasnikar & Grigoric 2002)

Despite the importance of Britain in the development of the cooperative movement and rapid growth of the number of cooperatives in the late twentieth century (Spear & Thomas 1997), the development of worker cooperatives has been relatively small-scale and concentrated in particular sectors. (Jones & Backus 1977; Spear & Thomas 1997; Estrin & Perotin 1989). Similar patterns are observed in North

America (Birchall 1997:215) In transitional (post-Soviet) republics, firms with high degrees of worker-ownership have been a fairly common product of privatization (Jones & Mygind 2000, Prasnikar and Gregoric 2002) and perform well, but often face government hostility that recalls that of the totalitarian states of the twentieth century (Birchall 1997:48,119) and may prove ephemeral. Sometimes cooperative property has been nationalized under the pretext of privatization! (Jones & Mygind 2000)

In Latin America, crisis conditions have led to the formation of worker cooperatives, including wild socializations in Argentina. (Fajn 2004, Birchall 1997:216) Here, again, it is difficult to gauge the permanence of these enterprises and thus their ultimate importance for the economy. Like many in Europe (Spear & Thomas 1997) and in the United States (McCain 1999) these enterprises are rescues of firms not viable as conventional firms. Cooperatives that emerge in this way inherit the problems of their predecessors and so may be relatively short-lived. In Latin America they also face highly changeable political conditions. In Brazil, in the last decades of the twentieth century, there was considerable development of worker cooperatives especially concentrated in professional activities. Cooperative medical practice has been particularly prominent, a response to the unemployment of doctors. (Bialosorski Neto 2001, Birchall 1997:217) In Africa and Asia, despite a few important experiments (Smith 2003a) there has been little development of autonomous worker cooperatives, as distinct both from state-organized compulsory collectives and craftsmen's cooperatives. (Birchall 1997:136, 157, 185)

Implications for Governance

The study of worker cooperatives remains a work in progress in both theoretical and

empirical aspects. Important work remains to be done, both in extending careful econometric studies to countries and industrial sectors that have been less studied and in clarification of theory. Nevertheless, the theoretical and empirical studies will support conclusions relevant for the study of public policy and governance.

We first consider implications for governance and policy at the national level and higher. A wise public policy will not prevent the owners of conventional firms from introducing employee participation, and may well find means to encourage it, since these measures support increased labor productivity. Moreover, the fact that they school their members in democratic processes in their most crucial daily activity will also recommend them to those who favor democracy.

Worker cooperatives have some advantages over other enterprise forms that public policies might support. Here again their democratic nature will seem an advantage to some. The tendency toward higher factor productivity could also be grounds for public support of this enterprise form, and could complement a wide variety of policy objectives by releasing resources for other uses. If the findings of Craig and Pencavel (1992) that the plywood cooperatives reduce pay rather than employment, can be generalized to other cooperatives, then promotion of worker cooperatives might favor macroeconomic stabilization. The major shortcomings of worker cooperatives would appear to be their relative difficulty in raising capital. Recognizing that enterprises of all kinds face liquidity constraints, it does appear that worker cooperatives are most affected by this form of market failure (and perhaps state enterprises least so.) This also makes it difficult for worker cooperatives to grow large enough to capture economies of scale.

As we have seen, these problems can be mitigated both by appropriate regulation and by federative networks of cooperatives, i.e. cooperation among cooperatives. The networks may be able to capture some economies of scale, (Smith 2003a:185-187, 2003b:207-209) and may be able to mitigate the liquidity constraint through mutual finance, for all of which the MCC provides excellent models. In the absence of a cooperative network of cooperatives, Smith's network externalities may to some extent be overcome, as in the case of U.S. plywood cooperatives. However, that case suggests that the cooperatives are likely to be concentrated in particular industries and the opportunities for complementarities in vertical integration not realized. The examples of Mondragon and La Lega are a useful contrast as, in these cases, the cooperative federations have facilitated vertical integration of complementary forms of production. Thus, a positive public policy in support of worker cooperatives should no less be a policy in favor of cooperative federations.

Regulation of worker cooperatives must be seen in a somewhat different light than the regulation of ownership-based conventional enterprises. An essential step is the provision of a legal form for the organization of cooperatives, which does not exist, for example, in Denmark (Mygind 1987) or in some American states. But in doing so, the state has perforce to define the sort of enterprise that is being recognized, and the line between definition and regulation is unclear. To some extent that is true also of conventional *corporations*, as some recent scandals illustrate; but the variety of existing cooperatives and the gap between the ideal and actual cooperatives pose deeper questions of definition.

- A worker cooperatives statute might, for example, follow the French example and

require open membership. Research on French and Italian cooperative groups indicates that this would ameliorate some of the shortcomings of worker cooperatives without surrendering the advantages of that form, so this seems an appropriate standard.

- The share price to become a member could have a fixed or market-determined price. Market-determined share prices have some advantages in encouraging a long investment horizon for individual workers, but the markets for these shares have been thin and not very efficient (Craig & Pencavel 1992), while Kamshad's open membership model suggests that the nominal price of a membership share is a passive variable in equilibrium if membership is open. A high price to buy into membership, which would require substantial borrowing on the part of liquidity-constrained employees, would conflict with the objective of open membership. All in all, the French model seems the better one here, too.
- The statute might require cooperatives to commit retained earnings out of profits, or some portion of total profits, to a reserve fund that cannot be recovered by the members in case the cooperative is dissolved. While Doucouliagos' meta-analysis (1995) suggests that this will somewhat reduce the productivity advantage of worker cooperatives, Kamshad (1997) argues that it will discourage the dissolution of successful cooperatives to sell them out to conventional firms. Thus, this issue is less clear and more research or experimentation may be needed.

All in all, the design of an appropriate cooperatives statute seems unavoidably a work of social engineering, but one that could

be very useful if firmly based in research results.

The research also has implications for the governance of the worker cooperatives themselves. Here we need to distinguish between the constitution and ongoing operation of a cooperative. At the founding, the prospective members are forming their social contract, and, even if not truly behind a veil of ignorance, are nevertheless relatively free to choose its constitutional form. Here, the research seems to indicate that a cooperative will probably be more successful if its constitution provides open membership and some lower limit on investment out of retained earnings. It seems clear that membership in a cooperative federation, if one is available, is to be recommended. A highly democratic organization, in which most or all operating authority rests with the general assembly of the members, is not inconsistent with a tight management in which daily operation is very much controlled by an individual manager; but regular consultation is likely to enhance labor productivity and thus member benefits. Supervision will be needed in the common interest, but needs consume relatively little of the cooperative's resources, as the cooperative can rely to a considerable extent on the reciprocity and mutual supervision of the members.

For the owners of conventional firms, the research results indicate that the *owners'* interests may be advanced by the introduction of participation in management decisions and profit-sharing, although executives dislike worker participation and an offset may be some difficulty in recruiting the best executives.

Finally, we may consider the implications for the governance of cooperative leagues and support bureaux. The research indicates that the most effective bodies are autonomous and highly democratic leagues focusing on worker-cooperatives. These bodies can be

effective and stabilizing, especially when they provide services that capture economies of scale, and when they facilitate vertical integration and complementarity among the members.

In all, while there is much to be learned, there seems to be a great potential for worker cooperatives and participatory enterprises to contribute to the improvement of governance.

Selected References

- Bakaikoa, Baleren; Anjel Errasti; and Agurtzane Begirstain. (2004) "Governance of the Mondragon Corporacion Cooperativa", *Annals of Public and Cooperative Economics*, 75, 1, 61-87.
- Bartlett, Will, John Cable, Saul Estrin, Derek Jones, and Stephen Smith. (1992) "Labor-managed cooperatives and private firms in North Central Italy: An empirical Comparison", *Industrial and Labor Relations Review*, 46, 1, 103-118.
- Battaille-Chedotel, Frederique and France Huntzinger. (2004) "Faces of Governance of Production Cooperatives: An Exploratory Study of Ten French Cooperatives", *Annals of Public and Cooperative Economics*, 75, 1, 89-111.
- Birchall, Johnston. (1997) *The International Co-Operative Movement*. Manchester University Press & St Martins Press.
- Baumol, William. (1953) "Firms with Limited Money Capital", *Kyklos*, 6, 2, 119-131.
- Bellas, Carl. (1972) *Industrial Democracy and the Worker-Owned Firm: A Study of Twenty-One Plywood Companies in the Pacific Northwest*. New York: Praeger.
- Berman, K.V. (1967) *Worker-Owned Plywood Companies*. Pullman, Washington: Washington State University.
- Bernstein, Paul. (1976) *Workplace Democratization: Its Internal Dynamics*. Kent, Ohio: Kent State University Press.

- Bialosorski Neto, Sigismundo. (2001) "Cooperative Development: Changes in the Brazilian Social Economy and Institutional Development", *Review of International Cooperation*, 94, 1, 59-65.
- Blasi, Joseph, Michael Conte, Douglas Kruse. (1996) "Employee Ownership and Corporate Performance Among Public Corporations", *Industrial and Labor Relations Review*, 50, 1, 60-79.
- Blasi, Joseph and Douglas Kruse. (1991) *The New Owners: The Mass Emergence of Employee Ownership in Public Companies and What it Means to American Business*. New York: Harper Business.
- Bonin, J.P., D.C. Jones, and L. Putterman. (1993) "Theoretical and Empirical Studies of Producer Cooperatives—Will Ever the Twain Meet", *Journal of Economic Literature*, 31, 3, 1290-1320.
- Bonin, John. (1981) "The Theory of the Labor-Managed Firm from the Membership's Perspective, with Implications for Marshallian Industry Supply", *Journal of Comparative Economics*, 5, 4, 337-352.
- Cable, John and F. Fitzroy. (1980) "Productivity, Efficiency, Incentives and Employee Participation", *Kyklos*, 33, 100-121.
- Carson, Robert. (1977) "A Theory of Cooperatives", *Canadian Journal of Economics*, 20, 4, 565-589.
- Chaves, Rafael. (1998) "Two Decades of Improvement of Spanish Worker Coops", *ICA Review*, 91, 1, 9-17.
- Craig, Ben and John Pencavel. (1992) "The Behavior of Worker Cooperatives: The Plywood Companies of the Pacific Northwest", *American Economic Review*, 82.5, 1083-1105.
- Craig, Ben and John Pencavel (1993) "The Objectives of Worker Cooperatives", *Journal of Comparative Economics* 17, 288-308.
- Craig, Ben and John Pencavel (1994) "The Empirical Performance of Orthodox Models of the Firm: conventional firms and worker cooperatives", *Journal of Political Economy* 104.4, 718-744.
- Craig, Ben and John Pencavel (1995) "Participation and productivity: A comparison of worker cooperatives and conventional firms in the plywood industry", *Brookings Papers on Economic Activity*, 121ff.
- Defourney, J.S., Saul Estrin, and Derek Jones (1985) "The Effects of Workers' Participation on Enterprise Performance: Empirical Evidence from French Cooperatives", *International Journal of Industrial Organization* 3, 197-217.
- Defourny, J.S. (1992) "Comparative Measures of Technical Efficiency for Five Hundred French Workers' Cooperatives", *Advances in the Economic Analysis of Participatory and Labor Managed Firms* ed. Derek Jones and Jan Svejnar (Greenwich, Conn: JAI Press) 27-62.
- Domar, Evsey (1966) "The Soviet Collective Farm as a Producer Cooperative", *American Economic Review* 56.4, 734-757.
- Doucouliaos, Chris (1995) "Worker Participation and Productivity in Labor-Managed and Participatory Capitalist Firms: A Meta-analysis", *Industrial and Labor Relations Review* 49.1, 58-77.
- Dow, Gregory K. and Putterman, Louis (2000) "Why Capital Suppliers (Usually) Hire Workers: What We Know and What We Need to Know", *Journal of Economic Behavior and Organization* 43.3, 319-336.
- Eichberger, Juergen (1989) "A Note on Bankruptcy Rules and Credit Constraints in Temporary Equilibrium", *Econometrica*, 57, 3, 707-715.
- Ellerman, David P. (1990) *The Democratic Worker-Owned Firm: A New Model for the East and the West*. Boston, London, Sydney, & Wellington: Unwin Hyman.

- Ely, Richard T. (1901) *An Introduction to Political Economy*. Revised Edition. New York: Eaton and Mains.
- Engberg, L. (1993) "Financing Employee-Managed Firms—Some Problems of a Wider Extension", *Economic and Industrial Democracy*, 14, 2.
- Errasti, Anjel Mari, Inaki Heras, Baleren Bakaikoa (2003) "The Internationalisation of Cooperatives: The Case of the Mondragon Cooperative Corporation", *Annals of Public and Cooperative Economics* 74.4, 553-584.
- Estrin, Saul and Virginie Perotin. (1987) "Producer Cooperatives: The British Experience", *International Review of Applied Economics*, 1, 2, 153-174.
- Estrin, Saul and Derek Jones. (1992) "The Viability of Employee-Owned Firms: Evidence from France", *Industrial & Labor Relations Review*, 45, 2, 323-338.
- Fajn, Gabriel. (2004) "Companies in Crisis -- Worker's Self Management", *Review of International Cooperation*, 97, 1, 93-100.
- Fehr, E. and Urs Fischbacher. (2004) "Third-party punishment and social norms", *Evolution and Human Behavior*, 25, 63-87.
- Flavin, M. (1981) "The Adjustment of Consumption to Changing Expectations About Future Income", *Journal of Political Economy*, 89, 5, 974-1009.
- Furubotn, Eirik and S. Pejovich. (1970) "Property Rights and the Behavior of the Firm in a Socialist State: The Example of Yugoslavia", *Zeitschrift fuer Nationaloekonomie*, 30, 3, 431-454.
- Furubotn, Eirik. (1976) "The Long-Run Analysis of the Labor-Managed Firm: An Alternative Interpretation", *American Economic Review*, 66, 1, 104-124.
- Furubotn, Eirik. (1980) "The Socialist Labor-Managed Firm and Bank-Finance Investment: Some Theoretical Issues", *Journal of Comparative Economics*, 4, 2, 184-191.
- Gui, Benedetto. (1985) "Limits to External Financing: A Model and an Application to Labor-Managed Firms", in D. Jones & J. Svejnar (Editors), *Advances in the Economic Analysis of Participatory and Labor-Managed Firms*, 107-120.
- Gunn, C. (1992) "Plywood Cooperatives in the United States—An Endangered Species", *Economic and Industrial Democracy*, 13, 4, 525-534.
- Hart, Oliver. (2000) *Different Approaches to Bankruptcy*. Working Paper 7921. New York: NBER.
- Hellwig, Martin F. (1981) "Bankruptcy, Limited Liability, and the Modigliani-Miller Theorem", *American Economic Review*, 71, 1, 155-170.
- International Cooperative Alliance. (1995) *Statement on the Co-Operative Identity*. www.ica.coop/coop/principles.html
- Jensen, Michael and William Meckling. (1979) "Rights and Production Functions: An Application to Labor-Managed Firms and Codetermination", *Journal of Business*, 52, 4, 469-506.
- Jones, Benjamin. (1968) *Co-Operative Production*. Oxford: Clarendon Press, repr. New York: Augustus Kelley.
- Jones, Derek and D. Backus. (1977) "British Producer Cooperatives in the Footwear Industry: An Empirical Evaluation of the Theory of Financing", *Economic Journal*, 87, 488-510.
- Jones, Derek. (1985) "The Cooperative Sector and Dualism in Command Economies: Theory and Evidence for the Case of Poland", *Advances in the Economics of Participatory and Labor-Managed Firms*, 1, 195-218.
- Jones, Derek and Niels Mygind. (2000) "The Effects of Privatization on Productive Efficiency: Evidence from the Baltic Republics", *Annals of Public and Cooperative Economics*, 71, 3, 415-439.

- Kahana, Nava and Shmuel Nitzan. (1993) "The Theory of the Labor-Managed Firm Revisited: The Voluntary-Interactive Approach", *Economic Journal*, 103, 419, 937-945.
- Kalecki, Michal. (1937) "The Principle of Increasing Risk", *Economica*, N.S. 4, 16, 440-447.
- Kamshad, Kimya. (1997) "A Model of the Free-Entry Produceer Cooperative", *Annals of Public and Cooperative Economics*, 68, 2, 225-245.
- Kelso, Louis O. and Patricia Hetter. (1967) *Two-Factor Theory: The Economics of Reality*. New York: Vintage Books.
- Kelso, Louis O. and Mortimer Adler. (1958) *The Capitalist Manifesto*. New York: Random House.
- Lutz, Mark and Kenneth Lux. (1979) *The Challenge of Humanistic Economics*. Menlo Park, CA: Benjamin and Cumming Publishers.
- Lutz, M.A. (1997) "The Mondragon Cooperative Complex: Community Enterprise in Action", *International Journal of Social Economics*, 24, 1404-1421.
- Major, Guy. (1996) "Solving the Underinvestment and Degeneration Problems of Workers' Cooperatives", *Annals of Public and Cooperative Economics*, 545-601.
- McCabe, Kevin, Stephen Rassenti and Vernon Smith. (1996) "Game Theory and Reciprocity in Some Extensive Form Games", *Proceedings of the National Academy of Science*.
- McCain, Roger A. (1977) "On the Optimum Financial Environment for Worker-Cooperatives", *Zeitschrift fuer Nationaloekonomie*, 37, 3/4, 355-384.
- McCain, Roger A. (1980) "A Theory of Codetermination", *Zeitschrift fuer Nationaloekonomie*, 40, 12, 65-90.
- McCain, Roger A. (1999) "The Mystery of Worker Buyouts of Bankrupt Firms", *Economic Analysis*, 2, 3, 165-186.
- McCain, Roger A. (2007) "Cooperation and Effort: Reciprocity and Mutual Supervision in Worker Cooperatives", in Sonja Novkovic (Editor), *Advances in the Economic Analysis of Participatory and Labor-Managed Firms*. Greenwich, Conn: JAI Press.
- Mill, John Stuart. (1987) *Principles of Political Economy*. A.M. Kelley, Reprint of 1909 edition.
- Mondragon Corporacion Cooperativa. (2001) *The History of an Experience*. www.mondragon.mcc.es/ing/quienessomos/historiaMCC_ing.pdf
- Mygind, Niels. (1987) "Are Self-Managed Firms Efficient? The Experience of Danish Fully and Partly Self-Managed Firms", in D. Jones and J. Svejnar (Editors), *Advances in the Economics of Participatory and Self-Managed Firms*. Greenwich, Conn: JAI Press) 243-323.
- Nishizumu, M. and J. Page. (1982) "Total Factor Productivity Growth, Technical Progress and Efficiency Change: Dimensions of Productivity Change in Yugoslavia", *Economic Journal*, 92, 920-936.
- Nozick, Robert. (1974) *Anarchy, State & Utopia*. New York: Basic Books.
- Perotin, Virginie. (1987) "Conditions of Survival and Closure of French Worker Cooperatives: Some Preliminary Findings", in D.C. Jones and J. Svejnar (Editors), *Advances in the Economic Analysis of Participatory and Labor-Managed Firms*. Greenwich, Conn: JAI Press, 201-224.
- Prasnikar, Janez and Jan Svejnar. (1988) "Economic Behavior of Yugoslav Enterprises", in Derek Jones and Jan Svejnar (Editors), *Advances in the Economic Analysis of Participatory and*

- Labor Managed Firms*. Greenwich, Conn: JAI Press, 237-311.
- Prasnikar, Janez and Aleksandra Gregoric. (2002) "The Influence of Workers' Participation on the Power of Management in Transitional Countries: The Case of Slovenia", *Annals of Public and Cooperative Economics*, 73, 2, 269-297.
- Robinson, Joan. (1967) "The Soviet Collective Farm as a Producer Cooperative: Comment", *American Economic Review*, 57, 1, 222-223.
- Romero, Antonio and Miguel Perez. (2003) "Organizational Culture, Individual Differences, and the Participation System in Cooperativism of Associated Workers in Andalusia, Spain", *Annals of Public and Cooperative Economics*, 74, 2, 283-320.
- Smith, Stephen. (2003a) "Network Externalities and Co-operative Networks: Stylized Facts and Theory", in Laixiang Sun (Editor), *Ownership and Governance of Enterprises*. Basingstoke, Hampshire, UK: Palgrave/Macmillan, 181-201.
- Smith, Stephen. (2003b) "Network Externalities and Co-operative Networks: A Comparative Case Study of Mondragon and La Lega with Implications for Developing and Transition Countries", in Laixiang Sun (Editor), *Ownership and Governance of Enterprises*. Basingstoke, Hampshire, UK: Palgrave/Macmillan, 202-241.
- Spear, Roger and Alan Thomas. (1997) "Comparative Perspective on Worker Cooperative Development in Several European Countries", *Annals of Public and Cooperative Economics*, 68, 3, 453-467.
- Stiglitz, Joseph. (1969) "A Re-Examination of the Modigliani-Miller Theorem", *American Economic Review*, 59, 5, 784-793.
- Thomas, H. and C. Logan (Editors), (1982) *Mondragon: An Economic Analysis*. London: Allen and Unwin.
- Vanek, Jaroslav. (1970) *The General Theory of Labor-Managed Market Economies*. Ithaca: Cornell University Press.
- Vanek, Jaroslav. (1971) *The Participatory Economy*. Ithaca: Cornell University Press.
- Vanek, Jaroslav. (1977) *The Labor-Managed Economy: Essays*. Ithaca and London: Cornell University Press.
- Ward, Benjamin. (1958) "The Firm in Illyria: Market Syndicalism", *American Economic Review*, 48, 3, 566-589.
- Whyte, W.F. and K.K. Whyte. (1988) *Making Mondragon: The Growth and Dynamics of the Mondragon Cooperative Complex*. Revised Edition. Ithaca, New York: Cornell University.

Roger Ashton McCain
 Department of Economics,
 Drexel University,
 Philadelphia, USA
 mccainra@drexel.edu

World Trade Organization and Environment

Eric Neumayer

Introduction

Environmentalists have mainly two concerns about the impact of the World Trade Organization (WTO) and its trade rules on the environment (Sampson and Chambers 2001; Neumayer 2001, 2004). The first one is that the multilateral trade regime as codified by the rules of the General Agreement on Tariffs and Trade (GATT) and WTO is, in their perspective at least, insensitive to environmental concerns. The enactment of strong environmental policies is rendered impossible as they would clash with countries' free trade obligations. Environmentalists see evidence for this in the relevant decisions GATT and WTO panels and appellate bodies have taken in disputes where environmental interests seemingly clashed with free trade interests. The second concern is that trade measures or substantive provisions contained in Multilateral Environmental Agreements (MEAs) might clash with WTO rules. They might therefore be the object of a potential future WTO dispute and might be judged inconsistent with a country's trade obligations. Also, the fear of such a future trade dispute might have a deterring effect on ongoing and future negotiations of MEAs to introduce trade measures or other substantive provisions if these are at the risk of being found inconsistent with WTO rules.

We first describe the origin and development of the GATT and WTO and the establishment of the Committee on Trade and Environment (CTE). We then present the basic rules governing the trade and environment relationship at the WTO. Following the two major concerns, we then discuss environmentally relevant

GATT/WTO panel and appellate body decisions and address the question whether trade measures and other provisions of MEAs clash with WTO rules. We conclude with a discussion of the future of the environment at the WTO.

Background: The Origin of the WTO

The origin of GATT/WTO stems from the failure to create a more ambitious organisation, namely the International Trade Organization (ITO). It was signed by 23 countries in 1947 and came into force in January 1948. Over time, more and more countries became contracting parties and several GATT trade rounds further decreased tariffs and extended the scope and reach of trade rules. The Uruguay Round, negotiated between 1986 and 1994, brought about the most far reaching changes. Rules on non-tariff measures were strengthened and new rules were established for new areas such as services, intellectual property, textiles, agriculture, which had not been subjected to multilateral rules before. The Uruguay Round also strengthened the existing GATT dispute settlement mechanism. Most importantly, however, it led to the creation of the World Trade Organization (WTO). The WTO finally gave a formal organisational structure to the multilateral trade regime. The GATT, as a legal text, still exists in a revised form as one of the WTO agreements (and arguably still its most important one). Therefore, while the GATT as an institution was superseded by the WTO, the GATT as an agreement establishing trade rules still exists.

Committee on Trade & Environment (CTE)

Already in 1971, that is even before the United Nations Conference on the Human Environment took place in Stockholm in 1972, GATT contracting parties agreed to establish a so-called Group on Environmental Measures and International Trade. However,

as the Group was supposed to convene only upon request and no GATT party submitted such a request, it lay dormant for twenty years. Finally, in February 1991 the GATT Director-General Arthur Dunkel convened the Group in order to inquire into the trade–environment nexus and make a contribution to the 1992 United Nations Conference on Environment and Development (UNCED) in Rio de Janeiro. Consequently, the Group met several times between November 1991 and January 1994 to discuss the consistency of trade provisions in multilateral environmental agreements (MEAs) with GATT rules, the trade effects of national environmental regulations and the trade effects of product packaging and eco-labelling rules (see WTO 1999, Annex 1).

At their Ministerial Meeting in Marrakesh in April 1994 trade ministers decided to widen and intensify the debates on the trade and environment linkage and to request the next meeting of the General Council of the WTO to establish a Committee on Trade and Environment (CTE) for that purpose. The Marrakesh Decision set up a list of ten items to be examined by the CTE. This list encompassed all the major areas of the international trading system including goods, services and intellectual property, the relationship between GATT rules including its dispute settlement mechanisms and MEAs and their dispute settlement mechanisms as well as market access for developing countries and arrangement for relations with non-governmental organisations (NGOs) and the transparency of WTO documentation.

The CTE met for the first time on 31 January 1995 and has since then held regular meetings. The results of these meetings have been rather disappointing in the eyes of environmentalists as no conclusive and definite results have emerged so far. Thus, the CTE reports usually tend to list the disagreements among GATT parties

concerning the items on their agenda, which, for lack of consensus, agree that nothing should be changed and the issues should be subject to further inquiry. The CTE has not turned out to be a frontrunner in triggering reform of the multilateral trade regime to make it more environmentally friendly, as hoped for by environmentalists.

Environmental Provisions of WTO Agreements

The preamble to the WTO Agreements states that its members are ‘seeking both to protect and preserve the environment and to enhance the means for doing so in a manner consistent with their respective needs and concerns at different levels of economic development’. The preamble is taken into account in interpreting WTO rules, but it does not have the same legal status as the substantive rights and obligations of WTO members codified in the GATT articles and other WTO rules themselves. The substantive WTO provisions most relevant to the environment are as follows:

- Article I (*Most favoured nation treatment*), which grants, with exceptions, most-favoured nation treatment to all members of the WTO. This means that any trade advantage given to any one WTO member must immediately and unconditionally granted to all other WTO members as well. In other words, *like* products must in principle be treated exactly the same independently from which trading partner the products originate as long as they are WTO members. It forbids discrimination between trading partners. Most-favoured nation treatment can be in conflict with MEAs as they often employ some form of trade measure against non-parties.
- Article III (*National treatment*), which goes one step beyond Art. I in not only forbidding discrimination between various like foreign products, but also forbidding

discrimination of foreign products relative to domestic ones. It grants, with exceptions, national treatment to foreign products. Of course, national treatment only applies to foreign products after having entered the market. A tariff on imported products does not constitute a violation of national treatment. Art. III is of direct environmental relevance to MEAs. Often they employ some distinction of products according to their process and production methods (PPMs), that is, according to the way in which these products were produced. The Montreal Protocol, for example, distinguishes between products manufactured with and without ozone depleting substances. The WTO rules by and large do not allow a distinction of products according to their PPMs as far as differences in environmental standards are concerned.

- Art. XI (*General elimination of quantitative restrictions*) is also of importance. Its objective is the general elimination of quantitative restrictions ‘other than duties, taxes or other charges, whether made effective through quotas, import or export licences or other measures’ (Art. XI:1). Environmentally motivated import restrictions that violate the national treatment provisions as well as export restrictions can be challenged as prohibited quantitative restrictions if they take the form of bans, embargoes and other prohibitions of trade.

Violations of these and other GATT provisions can be justified, however, with recourse to one of the ‘*General Exceptions*’ to the otherwise binding obligations of WTO members found in Article XX of GATT. Such exceptions are allowed if they are ‘necessary to protect human, animal or plant life or health’ or relate to ‘the conservation of exhaustible natural resources’. General

environmental protection is not listed, which is not surprising given that the Article has been drafted in 1947. However, dispute panels have tended to define these terms broadly and in the light of modern thinking about environmental policy (Neumayer 2001). Art. XX provides only a ‘limited and conditional exception from obligations’ (WTO 1998, p. 2), it is not a general escape clause. Importantly, the burden of proof lies with the party invoking one of the exceptions of Art. XX. That is, it is on the party that introduced a disputed measure to prove that in case other GATT articles are violated, the measure may nevertheless be justified under one of the exceptions of Art. XX (ibid, p. 3). Dispute panels have to decide first whether the disputed measure is covered by one of the exceptions and, if so, then proceed to assess whether it also satisfies the requirements of the preamble to Art. XX (ibid, p. 4). The other WTO agreements contain similar provisions – see Wiers (2002) and Charnovitz (2002) for a comprehensive overview.

Environmentally Relevant Dispute Settlement at the WTO

The WTO as the multilateral trade regime and its dispute settlement system is regarded as hindering countries from enacting strong environmental policies if they conflict with free trade. In the environmentalists’ view at least, the WTO system is inherently biased against environmental protection, giving free trade priority over the environment. As Friends of the Earth International (FoE 1999, p. 8) has put it: ‘Governments are increasingly using (or threatening to use) the WTO to challenge legitimate existing and proposed domestic and international laws as “barriers to trade”. They are able to do this because the WTO prioritises trade above all other societal values’.

Such argumentation is founded on early GATT and WTO dispute settlement, which

was arguably prioritising trade over environmental considerations (Neumayer 2001, 2004). However, more recent dispute settlement suggests that environmental considerations are given higher priority at the WTO. The most important recent environmentally relevant disputes are as follows:

- European Communities – Measures concerning meat and meat products (hormones). The US and Canada had challenged the import ban on beef from cattle raised with growth hormones.
- United States – Import prohibition of certain shrimp and shrimp products. The US had prohibited the importation of these products from countries that were not certified by the US as employing harvesting methods that prevented the incidental killing of five species of sea turtles. India, Pakistan, Thailand and Malaysia brought the dispute to the WTO.
- European Communities – Measures affecting asbestos and products containing asbestos. Canada had challenged a French and later European-wide ban of such products.

A detailed evaluation of these and other disputes can be found in Neumayer (2001, 2004). As a general trend, WTO dispute settlement has become increasingly accommodating to trade restrictions based on environmental grounds. It has defeated Canada's claim against the European Communities' (EC) restrictions of asbestos use and it has upheld US import restrictions on shrimp products from countries where shrimp harvesting accidentally kills sea turtles. It has decided against the EC restrictions of beef from cows raised with growth hormones, but it has done so because the EC had not provided a scientific assessment of the risks involved from growth hormones to human health. Once a scientific assessment demonstrates a plausible threat to

human health, a future WTO dispute panel might well uphold the ban. Future decisions will show whether this increased environmental friendliness of WTO dispute settlement is only a temporary reaction to the mounting pressure put on the WTO by public opinion and environmental lobby groups or a more permanent accommodation of environmental interests into the multilateral trade regime. In February 2006, a WTO body issued an interim ruling that found several aspects of the European Union's (EU) approval process for genetically modified organisms (GMOs) to be inconsistent with WTO rules, but the appellate body will have the final say on this highly contested issue.

WTO Rules and Multilateral Environmental Agreements

There is little doubt that WTO rules potentially clash with trade measures contained in many MEAs. However, since at the time of writing no WTO member had ever challenged any trade measure another WTO member had purportedly undertaken in compliance with an MEA, no relevant WTO case law and no binding interpretation exists – as of yet. Most MEAs with explicitly mandated or allowed for trade provisions restrict trade between parties and non-parties or even trade between parties. These restrictions certainly violate the general most favoured nation treatment obligation in GATT Art. I. If these restrictions take the form of import or export bans, export certificates or access restrictions rather than duties, taxes or other charges then they might violate the general elimination of quantitative restrictions obligation in GATT Art. XI. If countries in alleged pursuance to or compliance with MEAs applied regulations or taxes differently to imported than to domestically produced goods and services, then they might also violate their national

treatment obligation contained in GATT Art. III.

The MEAs with the greatest potential for a clash with WTO rules consist of the Montreal Protocol on Substances that Deplete the Ozone Layer, the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES), the Basel Convention on the Control of Transboundary Movements of Hazardous Wastes and Their Disposal, the Agreement on Persistent Organic Pollutants and the Kyoto Protocol and its potential follow-up conventions.

The potential for clash of WTO rules with MEAs could be dealt with in a number of different ways:

- Wait and see: The perhaps easiest way to deal with the potential for conflict is to do nothing for the time being until a dispute over such issues is initiated and to only embark on policy reform if it turns out to be necessary in the future.
- Temporary waiver: Art. IX:3 of the Agreement Establishing the WTO provides for the temporal waiving of WTO obligations, which could be used to waive certain obligations with respect to trade provisions in MEAs. Waiving decisions should normally be taken by consensus, otherwise by a three fourths majority. A waiver must be temporary with a fixed date of termination and is allowed in exceptional circumstances only (Art. IX:4).
- Interpretative statement: According to Art. IX:2 of the Agreement Establishing the WTO, the Ministerial Conference and the General Council of the WTO 'have the exclusive authority to adopt interpretations of this Agreement and of the Multilateral Trade Agreements'. An interpretation becomes adopted if it gains the support of three-fourths of WTO members.
- Amendment to GATT: The most far-reaching option is to amend the GATT. A

proposal for amendment needs a two-thirds majority and has effect only for those Members that have accepted the amendment (Art. X:3 of the Agreement Establishing the WTO). The amendment would need ratification by the accepting countries.

Future of the Environment at the WTO

In November 2001, the WTO Ministerial Meeting in Doha adopted a negotiation programme for trade and environment issues. In particular, ministers agreed to negotiate the relationship between WTO rules and obligations contained in MEAs and to reduce tariffs and non-tariff barriers to environmental goods and services (WTO 2001, paragraph 31). The CTE was instructed to review the effect of environmental measures on market access and labelling requirements for environmental purposes (*ibid.*, paragraph 32). This rather unambitious agenda for the so-called Millennium or Doha Round of trade negotiations is due to a fundamental conflict between developed and developing country WTO members with respect to the need for greening the WTO. Practically all developed countries – partly by conviction, partly due to pressure from NGOs – are to some extent in favour of such greening. On the other hand, practically all developing countries are either strictly opposed to or at least most reluctant to accept even negotiation of such reform proposals. That the environment is on the trade negotiation agenda at all is due to the threat of developed countries and the EU in particular to refuse negotiation on other issues if the environment is left out. Developing country representatives do not trust the alleged idealistic intentions of the proponents (see DeSombre (1995) for evidence of protectionist motivations for environmental measures). Instead they regard the proposals as motivated by economically protectionist

reasons (Neumayer 2001). This hostility has its roots a much deeper frustration with the distribution of benefits in the multilateral trade regime. In the view of developing countries, the developed countries benefit much more from the WTO and its agreements than they themselves do. A greening of the WTO will therefore only be achievable if developing countries can be convinced that this will not run counter to their economic development aspirations.

Selected References

- Charnovitz, Steve. (2002) *Trade Law and Global Governance*, Cameron May, London.
- DeSombre, E.R. (1995) "Baptists and Bootleggers for the Environment: The Origins of United States Unilateral Sanctions", *Journal of Environment & Development*, 4, 53-75.
- FoE. (1999) *WTO Scorecard – WTO and Free Trade vs. Environment and public health: 4–0*, Friends of the Earth, Washington, DC.
- Neumayer, Eric. (2001) *Greening Trade and Investment—Environmental Protection without Protectionism*. London: Earthscan.
- Neumayer, Eric. (2004) "The WTO and the Environment: Its Past Record is Better than Critics Believe, but the Future Outlook is Bleak", *Global Environmental Politics*, 4, 1-8.
- Sampson, Gary and W. Bradnee Chambers (2001) (Editors) *Trade, Environment, and the Millennium*. Second Edition. Tokyo: United Nations University Press.
- Wiers, Joachim. (2002) *Trade and Environment in the EC and the WTO: A Legal Analysis*. Groningen: Europa Law Publishing.
- WTO (1998) *GATT/WTO Dispute Settlement Practice Relating to Article XX, Paragraphs (b) (d) and (g) of GATT*, Note

by the Secretariat, WT/CTE/W/53/Rev.1. Geneva: World Trade Organization.

WTO (1999) *Background Document for High Level Symposium on Trade and Environment*, Geneva, 15–16 March 1999. Geneva: World Trade Organization.

WTO (2001) *Doha Ministerial Declaration*, WT/MIN(01)/DEC/1, World Trade Organization, Geneva.

Websites

- World Trade Organisation. www.wto.org.
- World Trade Organisation. *For NGOs*. www.wto.org/english/forums_e/ngo_e/ngo_e.htm
- International Centre for trade and Sustainable Development. www.ictsd.org

Eric Neumayer
Geography and Environment Department
London School of Economics
London, UK
e.neumayer@lse.ac.uk